



IntechOpen

# Finite Element Method Applications

Bridging Mathematical Theory  
and Real-World Problem-Solving

*Edited by Mahboub Baccouch*





---

# Finite Element Method Applications - Bridging Mathematical Theory and Real-World Problem- Solving

*Edited by Mahboub Baccouch*

Published in London, United Kingdom

---

Finite Element Method Applications – Bridging Mathematical Theory and Real-World Problem-Solving  
<http://dx.doi.org/10.5772/intechopen.1009078>  
Edited by Mahboub Baccouch

#### Contributors

Aldo Sarmiento Guevara, Ana López-Palencia, Anderson Alvarenga de Moura Meneses, Babar Hayat, Carlos Reales-Martínez, Chuizhou Meng, Fabián A. Ortega-Quintana, Francisco Badea-Romero, Geoffrey Lossa, Imran Shah, Jesús A. Perez-Fernandez, Jorge L. Díaz Calle, Lenilson Moreira Araujo, Li Chen, Mahboub Baccouch, Olivier Deblecker, Peng Wang, Roger Mestas Chávez, Shijun Ji, Xiaoyu Cao, Xiufeng Li, Yu Lv

© The Editor(s) and the Author(s) 2026

The rights of the editor(s) and the author(s) have been asserted in accordance with the Copyright, Designs and Patents Act 1988. All rights to the book as a whole are reserved by INTECHOPEN LIMITED. The book as a whole (compilation) cannot be reproduced, distributed or used for commercial or non-commercial purposes without INTECHOPEN LIMITED's written permission. Enquiries concerning the use of the book should be directed to INTECHOPEN LIMITED rights and permissions department ([permissions@intechopen.com](mailto:permissions@intechopen.com)).

Violations are liable to prosecution under the governing Copyright Law.



Individual chapters of this publication are distributed under the terms of the Creative Commons Attribution 4.0 License which permits commercial use, distribution and reproduction of the individual chapters, provided the original author(s) and source publication are appropriately acknowledged. If so indicated, certain images may not be included under the Creative Commons license. In such cases users will need to obtain permission from the license holder to reproduce the material. More details and guidelines concerning content reuse and adaptation can be found at <http://www.intechopen.com/copyright-policy.html>.

#### Notice

Statements and opinions expressed in the chapters are those of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in London, United Kingdom, 2026 by IntechOpen  
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, 167-169 Great Portland Street, London, W1W 5PF, United Kingdom

For EU product safety concerns: IN TECH d.o.o., Prolaz Marije Krucifikse Kozulić 3, 51000 Rijeka, Croatia, [info@intechopen.com](mailto:info@intechopen.com) or visit our website at [intechopen.com](http://intechopen.com).

#### British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

Finite Element Method Applications – Bridging Mathematical Theory and Real-World Problem-Solving  
Edited by Mahboub Baccouch

p. cm.

Print ISBN 978-1-83635-932-6

Online ISBN 978-1-83635-931-9

eBook (PDF) ISBN 978-1-83635-933-3

If disposing of this product, please recycle the paper responsibly.

# Meet the editor



Mahboub Baccouch is a Professor of Mathematics at the University of Nebraska at Omaha. He earned his Ph.D. in Mathematics from Virginia Tech in 2008 and specializes in applied and computational mathematics, focusing on numerical methods for deterministic and stochastic differential equations, including finite element and discontinuous Galerkin methods. He has authored more than 75 peer-reviewed publications, two textbooks, four book chapters, and one encyclopedia article. His contributions have been recognized with the Isaacson Professorship, the James Earl Diamond Professorship in Mathematics, the College of Arts and Sciences Excellence in Research Award, and the Kerrigan Award for Excellence in Teaching. Dr. Baccouch is also dedicated to mentoring and guiding students in both research and education.



# Contents

|                                                                                                                                                                                                                                                                                               |           |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>Preface</b>                                                                                                                                                                                                                                                                                | <b>IX</b> |
| <b>Section 1</b>                                                                                                                                                                                                                                                                              |           |
| Foundations of the Finite Element Method                                                                                                                                                                                                                                                      | <b>1</b>  |
| <b>Chapter 1</b>                                                                                                                                                                                                                                                                              | <b>3</b>  |
| Overview of the Finite Element Method for a Class of Partial Differential Equations<br><i>by Mahboub Baccouch</i>                                                                                                                                                                             |           |
| <b>Section 2</b>                                                                                                                                                                                                                                                                              |           |
| Classical and Advanced FEM Formulations                                                                                                                                                                                                                                                       | <b>31</b> |
| <b>Chapter 2</b>                                                                                                                                                                                                                                                                              | <b>33</b> |
| The Classical Galerkin Finite Element Method Applied to Neutron Diffusion Theory<br><i>by Lenilson Moreira Araujo and Anderson Alvarenga de Moura Meneses</i>                                                                                                                                 |           |
| <b>Chapter 3</b>                                                                                                                                                                                                                                                                              | <b>53</b> |
| Numerical Analysis of the Enriched Mixed Method, Evaluating Computational Costs and Errors through Comparisons with Formulations in Continuous Functions, Discontinuous Functions, and Mixed Finite Elements<br><i>by Jorge L. Díaz Calle, Roger Mestas Chávez and Aldo Sarmiento Guevara</i> |           |
| <b>Section 3</b>                                                                                                                                                                                                                                                                              |           |
| Applications in Engineering and Science                                                                                                                                                                                                                                                       | <b>73</b> |
| <b>Chapter 4</b>                                                                                                                                                                                                                                                                              | <b>75</b> |
| FEM Modeling Techniques for Industrial Applications<br><i>by Francisco Badea-Romero, and Jesús A. Perez-Fernandez</i>                                                                                                                                                                         |           |
| <b>Chapter 5</b>                                                                                                                                                                                                                                                                              | <b>95</b> |
| Finite Element Method Applied to Heat Transfer in Food Processing<br><i>by Ana López-Palencia, Carlos Reales-Martínez and Fabián A. Ortega-Quintana</i>                                                                                                                                       |           |

|                                                                               |            |
|-------------------------------------------------------------------------------|------------|
| <b>Chapter 6</b>                                                              | <b>117</b> |
| Finite Element Analysis for Micromixing Applications                          |            |
| <i>by Imran Shah, Babar Hayat and Shijun Ji</i>                               |            |
| <b>Chapter 7</b>                                                              | <b>137</b> |
| Finite Element Method for Design of Flexible Tactile Sensors Toward           |            |
| Wearable Monitoring Applications                                              |            |
| <i>by Peng Wang, Yu Lv, Li Chen, Xiufeng Li, Xiaoyu Cao and Chuizhou Meng</i> |            |
| <b>Chapter 8</b>                                                              | <b>167</b> |
| Uncertainty Quantification in Electromagnetic Computation                     |            |
| <i>by Geoffrey Lossa and Olivier Deblecker</i>                                |            |

# Preface

The finite element method (abbreviated FEM) has emerged as one of the most versatile and widely used numerical techniques for the approximation of solutions to partial differential equations (PDEs) across science, engineering, and technology. This edited volume aims to provide both a comprehensive introduction to the fundamental principles of FEM and an exploration of its diverse applications, ranging from classical numerical formulations to cutting-edge industrial and biomedical implementations. The chapters are organized to guide the reader from foundational concepts to specialized applications, reflecting both theoretical rigor and practical relevance.

The volume begins with **Overview of the Finite Element Method for a Class of Partial Differential Equations**, which introduces the core ideas and framework of FEM. This chapter covers the historical development of the method, its mathematical foundations, and its application to key classes of PDEs, including second-order elliptic problems, the heat equation, the reaction-convection-diffusion equation, the wave equation, and the Navier-Stokes equations. This chapter establishes a solid conceptual basis for readers, preparing them for more advanced formulations and applications discussed in subsequent chapters.

In **The Classical Galerkin Finite Element Method Applied to the Neutron Diffusion Theory**, the reader is introduced to the Galerkin finite element method (GFEM) and its application to reactor physics. The chapter presents a step-by-step variational formulation of the neutron diffusion equation and demonstrates implementation using triangular and quadrilateral elements. Benchmark problems illustrate convergence properties and the practical accuracy of the classical FEM approach compared to finite difference methods.

The chapter **Numerical Analysis of the Enriched Mixed Method, Evaluating Computational Costs and Errors through Comparisons with Formulations in Continuous Functions, Discontinuous Functions, and Mixed Finite Elements** explores advanced FEM formulations for diffusion problems. Here, enriched mixed methods are compared with classical, continuous, and discontinuous FEM approaches, highlighting computational costs and approximation errors for both primal and dual variables. The chapter provides valuable insights into the optimization of FEM formulations, offering guidance for choosing methods that balance accuracy and computational efficiency.

Applications of FEM in real-world scenarios are explored in the next part of the volume. **FEM Modeling Techniques for Industrial Applications** presents practical strategies for engineering simulations, including element selection, meshing approaches, material modeling, and connection modeling. Emphasis is placed on techniques that achieve reliable results while minimizing computational cost, reflecting the practical constraints of industrial engineering practice.

**Finite Element Method Applied to Heat Transfer in Food Processing** illustrates the use of FEM for axisymmetric and multi-domain heat transfer problems in food engineering. The chapter demonstrates the use of open-source FEM software, such as FEniCS, Gmsh, and ParaView, to model blanching and cooling processes for various food products, providing readers with both methodological guidance and practical examples.

**Finite Element Analysis for Micromixing Applications** focuses on microfluidic systems, where efficient fluid mixing is essential for biomedical, chemical, and energy applications. The chapter presents FEA simulations of various micromixer designs, including Y-shaped split-and-recombine, Cantor fractal-based, and obstacle-based geometries. Comparative analyses highlight optimal design parameters, mixing performance, and practical implementation considerations.

In **Finite Element Method for Design of Flexible Tactile Sensors Toward Wearable Monitoring Applications**, FEM is applied to the design and optimization of flexible sensors for wearable electronics. The chapter emphasizes multiphysics coupling, constitutive modeling, and structural optimization strategies. Applications include piezoresistive, capacitive, triboelectric, and temperature sensors, showcasing how FEM can accelerate sensor development, reduce experimental cost, and support system-level integration.

Finally, **Uncertainty Quantification in Electromagnetic Computation** integrates probabilistic modeling with FEM to assess the sensitivity of electromagnetic simulations to uncertain input parameters. The chapter introduces techniques such as polynomial chaos expansion and Bayesian inference, providing readers with tools to quantify uncertainty and improve the robustness of numerical predictions.

This volume would not have been possible without the dedicated contributions of all chapter authors, whose expertise and diligence have produced clear and comprehensive expositions across theory, computation, and application. I am grateful to the Co-Editors and the editorial team for their support throughout the editorial process.

It is our hope that this book will serve as a valuable resource for students, researchers, and practitioners in applied mathematics, engineering, and computational sciences, providing both a solid foundation in FEM theory and a practical perspective on its diverse applications.

**Mahboub Baccouch**  
Department of Mathematical and Statistical Sciences,  
University of Nebraska at Omaha,  
Omaha, Nebraska, USA

---

## Section 1

# Foundations of the Finite Element Method

---



## Chapter 1

# Overview of the Finite Element Method for a Class of Partial Differential Equations

*Mahboub Baccouch*

### Abstract

The Finite Element Method (FEM) is a numerical framework commonly employed to approximate solutions of differential equations that describe complex physical and engineering phenomena. Originating in the 1950s as a tool for addressing difficult problems in structural mechanics and elasticity – particularly those governed by elliptic partial differential equations (PDEs) – it has since evolved into a general numerical framework applicable to a broad range of scientific and engineering disciplines. Today, FEM is employed in diverse areas such as electromagnetics, fluid dynamics, structural analysis, heat transfer, aerospace, biomechanics, automotive, computer graphics, biomedical research, manufacturing, and construction. This chapter offers a concise introduction to the fundamental principles of FEM, serving as the foundation for the topics developed in later chapters. Our goal is to present FEM as a unified approach for approximating the solutions of key classes of PDEs. In particular, we apply the method to second-order elliptic PDEs, the heat (or diffusion) equation, the reaction-convection-diffusion equation, the wave equation, and the Navier-Stokes equations.

**Keywords:** Finite Element Method, elliptic problems, error estimation, heat/diffusion equation, convection-diffusion equation, wave equation, Navier-Stokes equations

### 1. Introduction

Partial differential equations (PDEs) provide a fundamental framework for modeling real-world phenomena across numerous disciplines, including engineering, applied mathematics, the physical and life sciences, economics, and many other areas. They provide a natural framework for describing processes that evolve over time or space, such as heat transfer, fluid flow, population dynamics, chemical reactions, financial markets, and mechanical vibrations. Unfortunately, exact solutions to PDEs can seldom be obtained in closed form. In many practical situations, the equations are too complex – owing to nonlinearities, irregular geometries, variable coefficients, or complicated boundary conditions – to be solved

analytically. As a result, the development and use of numerical methods have become essential for approximating their solutions and making meaningful predictions.

A wide variety of numerical techniques have been designed for this purpose. Some of the most commonly employed include the finite difference method (FDM), which replaces derivatives with difference quotients defined on a computational grid; the finite element method (FEM), which subdivides the domain into elements and uses piecewise polynomial approximations; the finite volume method (FVM), which enforces conservation laws over control volumes; spectral methods, which employ global basis functions to achieve high accuracy for smooth problems; and the discontinuous Galerkin method (DGM), which unites the best features of the FEM and the FVM, offering exceptional flexibility in solving complex problems. These methods provide powerful tools for tackling differential equations that are analytically intractable, making it possible to approximate and study problems of practical importance across science, engineering, and industry. This chapter offers an introduction to the FEM, highlighting its use as a general framework for solving boundary value problems (BVPs). The main ideas are presented through simple model equations to build intuition for more advanced applications in later chapters.

The FEM has become the standard numerical scheme for approximating the solutions of many mathematical problems; see, for example, [1–11] and the references therein. The FEM is one of the most powerful and widely used numerical techniques for solving PDEs encountered in science and engineering. It provides a systematic framework for approximating solutions of BVPs through the use of local basis (shape) functions defined over a mesh partition of the computational domain. Through this process, the PDE is converted into a discrete algebraic system, whose solutions approximate the values of the unknown function at the mesh nodes. The method was initially developed to tackle difficult problems in structural analysis and elasticity, especially within the fields of civil and aeronautical engineering. Its early development can be traced to the pioneering work of Hrennikoff [12] and Courant [13], who introduced the idea of approximating variational problems using piecewise polynomial functions. Although Hrennikoff and Courant adopted different approaches (Hrennikoff through lattice models and Courant through variational principles), they both arrived at the same fundamental concept: dividing a continuous domain into smaller subdomains, now known as elements. In the 1950s, Turner, Clough, Martin, and Topp developed FEM as a practical computational tool for structural mechanics, laying the foundation for its broad adoption in engineering. The term *finite element method* was first used by Clough [14]. During the 1960s and 1970s, the method rapidly evolved both mathematically and computationally. The Dayton conferences (1965–1970) helped unify researchers in the field, while major advances such as isoparametric elements, numerical integration, and hybrid formulations were developed by Zienkiewicz, Irons, and Oden [15, 16]. A fundamental mathematical foundation for the FEM was later established by Gilbert Strang and George Fix [17], whose work provided rigorous justification for the method's effectiveness. Over time, the FEM has evolved and expanded into a versatile computational framework, now serving as a fundamental tool for simulating physical processes in many branches of engineering, such as fluid mechanics, heat conduction, and electromagnetics.

The mathematical theory of FEM was established during the 1960s and 1970s through the seminal contributions of Strang and Fix [17], Ciarlet [5], and Babuška

and Aziz [18]. These works provided rigorous error analysis, convergence theory, and stability results, giving the method a firm theoretical basis. The 1980s and 1990s witnessed further developments in FEM theory and applications. Mixed and hybrid finite element methods were developed and analyzed in the influential monograph by Brezzi and Fortin [19]. Adaptive finite element methods, driven by a posteriori error estimation, were introduced by Babuška and Rheinboldt [20] and further developed by Verfürth [21].

Since then, FEM has evolved into a general framework for the numerical approximation of a wide variety of BVPs. Over the years, numerous extensions and refinements have been developed. Mixed and hybrid finite element methods [19] were introduced to handle constraints such as incompressibility. Discontinuous Galerkin (DG) methods [22, 23] combine the flexibility of finite volume schemes with the accuracy of FEM. Adaptive finite element techniques [20, 21] have further improved computational efficiency by locally refining the mesh based on a posteriori error estimates.

Modern developments have expanded FEM into new domains, including high-order and isogeometric analysis [24], stochastic finite element methods for uncertainty quantification [25], and applications in computational biology, materials science, and finance. Comprehensive treatments of the finite element method can be found in standard references such as Strang and Fix [17], Ciarlet [5], Brenner and Scott [4], and Johnson [6].

In recent decades, FEM has continued to evolve in several directions. High-order and spectral element methods [26] have improved accuracy and efficiency for smooth problems. The advent of isogeometric analysis [24] has bridged the gap between computer-aided design (CAD) and analysis. Stochastic finite element methods [25] have been developed to quantify uncertainty in physical systems, while domain decomposition and parallel algorithms [27] have enabled large-scale computations. Modern treatments, such as those by Ern and Guermond [28] and Hughes [29], provide comprehensive overviews of both theoretical and computational aspects. Today, the finite element method stands as a mature and flexible computational framework applicable to a wide range of disciplines, from fluid dynamics and electromagnetics to biomechanics, materials science, and data-driven modeling. The continued development of adaptive, stochastic, and high-performance FEM algorithms ensures its central role in modern computational science and engineering.

The FEM offers several important advantages, which can be outlined as follows:

- (1) **Numerical efficiency:** Dividing the computational domain into finite elements produces system matrices that are typically sparse and symmetric. Consequently, the method enables highly efficient management of both storage requirements and the solution of the resulting algebraic system.
- (2) **Treatment of nonlinearities:** The FEM provides a robust framework for capturing nonlinear material behavior, such as stress-strain nonlinearity and hysteresis effects.
- (3) **Flexibility with complex geometries:** The method can efficiently handle domains of any geometry, using triangular elements for 2D problems and tetrahedral elements for 3D problems.

- (4) **Broad applicability:** The FE method is highly versatile and has been successfully applied in numerous fields, including structural mechanics, heat transfer, electromagnetics, fluid flow, acoustics, and multiphysics problems.

At the same time, the FEM has certain limitations, such as high computational cost and the need for careful mesh design, which make it particularly appropriate for specific classes of PDEs, like the ones presented in this chapter.

There are many widely used FEM software packages, including ANSYS, Abaqus, NASTRAN, and ADINA, each offering specialized capabilities for different engineering applications. Among these, COMSOL Multiphysics (formerly known as FEMLAB until 2005) stands out as a versatile tool for simulating a broad spectrum of physical phenomena. It has become a standard in both academia and industry for research, development, and design. COMSOL is particularly well-suited for multiphysics problems, where several coupled physical processes must be modeled simultaneously by numerically solving the underlying PDEs using the FEM. It supports one-, two-, and three-dimensional simulations and offers a modern graphical user interface that allows users to define geometries, boundary conditions, materials, and physics with ease. In addition, it provides flexible options for customization and automation, including integration with MATLAB and scripting through its native Java API, making it suitable for both interactive modeling and advanced computational workflows. One of COMSOL's key strengths lies in its ability to seamlessly couple different physical processes – such as thermal, structural, fluid, and electromagnetic phenomena – within a single simulation framework, enabling accurate modeling of complex real-world systems.

This chapter introduces the FEM through simple two-dimensional model problems. We begin with the formulation of the FE method, including the weak form, domain discretization, and assembly of the algebraic system. We also discuss basis functions, interpolation, shape functions, and numerical integration. Finally, we recall key error estimates and convergence properties, illustrating how mesh size and polynomial degree affect solution accuracy. Throughout, examples highlight the method's application to structural mechanics, heat conduction, fluid flow, and electromagnetic problems, providing a foundation for more advanced topics in later chapters. For further details, the reader is referred to classical textbooks such as [1, 7, 15, 29, and 30].

## 2. The FEM for elliptic PDEs

In this section, we present the application of the finite element method to an elliptic boundary value problem. Our aim is to determine a function  $u = u(x), x \in \Omega \subset \mathbb{R}^d, d = 2, 3$  that satisfies [1]

$$-\nabla \cdot (a(\mathbf{x})\nabla u) + b(\mathbf{x})u = f(\mathbf{x}), \quad \text{in } \Omega, \quad \mathbf{n} \cdot a\nabla u = \epsilon(g(\mathbf{x}) - u) + q(\mathbf{x}), \quad \text{on } \partial\Omega, \quad (1)$$

where the functions  $a > 0$  and  $b \geq 0$  are smooth, the parameter  $\epsilon \geq 0, f \in L^2(\Omega)$ , and  $g, q \in C(\partial\Omega)$ . Next, we express the equation in its variational form and seek an approximate solution within the space of continuous piecewise polynomials.

## 2.1 Meshes

Suppose the domain  $\Omega \subset \mathbb{R}^2$  is bounded and has a polygonal boundary  $\partial\Omega$ . A triangulation  $\mathcal{T}_h$  of the region  $\Omega$  is a collection of elements (or triangles)  $T$  satisfying  $\Omega = \cup_{T \in \mathcal{T}_h} T$ , where any two distinct elements intersect only at a shared edge  $E$ , a shared node vertex (or node), or not at all. The vertices of the triangles are referred to as *nodes*. For each triangle  $T$ , we define its diameter as  $h_T = \text{diam}(T)$ , typically defined by the length of the triangle's longest edge.

Suppose that the triangulation  $\mathcal{T}_h$  has  $N$  nodes and  $M$  triangles. The data can be represented on a computer in two matrices. The point matrix  $P \in \mathbb{R}^{2 \times N}$  contains the coordinates of the nodes  $N_i = (x_i, y_i)$  for  $i = 1, 2, \dots, N$ , while the connectivity matrix  $K \in \mathbb{R}^{3 \times M}$  encodes the triangles, indicating which vertices (numbered from 1 to  $N$ ) constitute each element  $T$  and the triangle's orientation. These matrices have the following forms

$$P = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \\ y_1 & y_2 & \cdots & y_N \end{bmatrix}, \quad K = \begin{bmatrix} n_1^\alpha & n_2^\alpha & \cdots & n_M^\alpha \\ n_1^\beta & n_2^\beta & \cdots & n_M^\beta \\ n_1^\gamma & n_2^\gamma & \cdots & n_M^\gamma \end{bmatrix}.$$

This implies that each triangle  $T_i$  in the mesh  $\mathcal{T}_h$  is defined by its three nodes  $n_i^\alpha$ ,  $n_i^\beta$ , and  $n_i^\gamma$ , numbered in a counterclockwise order.

MATLAB's built-in Delaunay function forms a triangulation in which the triangles are defined by the given points as their vertices. A Delaunay triangulation is regarded as optimal since it maximizes the smallest angle among all triangles, thereby preventing the formation of overly sharp angles.

MATLAB includes a built-in PDE Toolbox, which provides a mesh generation algorithm for creating finite element meshes. The toolbox supports both two- and three-dimensional domains, enabling the generation of triangular and tetrahedral meshes, with options for mesh refinement to enhance simulation accuracy.

## 2.2 Continuous piecewise polynomial spaces

Consider a triangle  $T$  in the mesh  $\mathcal{T}_h$  with vertices  $N_i = (x_i, y_i)$  for  $i = 1, 2, 3$ . Define

$$P^1(T) = \{v \in C(T) : v(x, y) = c_1 + c_2x + c_3y, (x, y) \in T, c_1, c_2, c_3 \in \mathbb{R}\},$$

to be the vector space of linear polynomials in two dimensions on  $T$ , where  $C(T)$  is used to denote the space of continuous functions on the triangle  $T$ . Define  $v_i = v(N_i)$ ,  $i = 1, 2, 3$ . Observe that any  $v \in P^1(T)$  is uniquely defined by the nodal values  $\{v_i\}_{i=1}^3$ . Indeed, once the nodal values are specified, the coefficients  $c_i$ ,  $i = 1, 2, 3$  can be found by solving the linear system of equations:

$$\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}.$$

We observe that the above system has a unique solution since

$$\det \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} = 2|T| \neq 0, \quad \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix}^{-1} = \frac{1}{2|T|} \begin{bmatrix} x_2y_3 - x_3y_2 & x_3y_1 - x_1y_3 & x_1y_2 - x_2y_1 \\ y_2 - y_3 & y_3 - y_1 & y_1 - y_2 \\ x_3 - x_2 & x_1 - x_3 & x_2 - x_1 \end{bmatrix},$$

where  $|T| = \frac{1}{2}(x_2y_3 + x_3y_1 + x_1y_2 - x_3y_2 - x_1y_3 - x_2y_1)$ , whose absolute value represents the area of  $T$ .

Define  $\lambda_i \in P^1(T)$  by the nodal conditions  $\lambda_i(N_j) = \delta_{ij}$ , where  $\delta_{ij}$  denotes the Kronecker delta symbol. With this definition, any  $v \in P^1(T)$  can be expressed as

$$v = \alpha_1\lambda_1 + \alpha_2\lambda_2 + \alpha_3\lambda_3,$$

where  $\alpha_i = v(N_i)$ ,  $i = 1, 2, 3$ . In order to find the basis functions  $\lambda_i$ , we assume it has the form

$$\lambda_i = a_i + b_ix + c_iy, \quad i = 1, 2, 3, \quad (x, y) \in T.$$

Applying the nodal conditions  $\lambda_i(N_j) = \delta_{ij}$  gives

$$B \begin{bmatrix} a_1 \\ b_1 \\ c_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad B \begin{bmatrix} a_2 \\ b_2 \\ c_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad B \begin{bmatrix} a_3 \\ b_3 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \text{where } B = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix}.$$

Upon solving the system, we obtain

$$\lambda_1 = \frac{1}{2|T|} \left( x_2y_3 - x_3y_2 + (y_2 - y_3)x + (x_3 - x_2)y \right),$$

$$\lambda_2 = \frac{1}{2|T|} \left( x_3y_1 - x_1y_3 + (y_3 - y_1)x + (x_1 - x_3)y \right),$$

$$\lambda_3 = \frac{1}{2|T|} \left( x_1y_2 - x_2y_1 + (y_1 - y_2)x + (x_2 - x_1)y \right).$$

If  $\mathcal{T}_h$  is a triangulation of the computational domain  $\Omega$ . Then we introduce the finite element space

$$V_h = \{v \in C(\Omega) : v|_T \in P^1(T), \forall T \in \mathcal{T}_h\}.$$

Thus, members of  $V_h$  are continuous and piecewise linear function. Moreover any  $v \in V_h$  is uniquely defined by its nodal values  $\{v(N_i) : i = 1, 2, \dots, N\}$ . Define the nodal basis functions  $\phi_j \in V_h$  by the condition  $\phi_j(N_i) = \delta_{ij}$  for  $i, j = 1, \dots, N$ . Then the set  $\{\phi_j\}_{j=1}^N$  forms a basis for  $V_h$  (the so-called hat functions). Consequently, any function  $v \in V_h$  is expressed as

$$v(x, y) = \sum_{i=1}^N \alpha_i \phi_i(x, y),$$

where, for each  $i = 1, 2, \dots, N$ ,  $\alpha_i = v(N_i)$ .

## 2.3 Interpolation

Let  $u$  be a continuous function defined on a triangle  $T \in \mathcal{T}_h$ . Suppose its nodes are  $N_i = (x_i, y_i)$  for  $i = 1, 2, 3$ . We define the interpolant of  $u$  by

$$\pi u(x, y) = \sum_{i=1}^3 u(N_i) \phi_i(x, y).$$

Note that  $\pi u(N_i) = u(N_i)$  for  $i = 1, 2, 3$ . To estimate the interpolation error, defined by  $u - \pi u$ , we first define the following norms:

$$\|u\|_{L^2(\Omega)}^2 = \int_{\Omega} |u|^2 dx dy, \quad \|Du\|_{L^2(\Omega)}^2 = \|u_x\|_{L^2(\Omega)}^2 + \|u_y\|_{L^2(\Omega)}^2,$$

$$\|D^2u\|_{L^2(\Omega)}^2 = \|u_{xx}\|_{L^2(\Omega)}^2 + 2\|u_{xy}\|_{L^2(\Omega)}^2 + \|u_{yy}\|_{L^2(\Omega)}^2.$$

The proof of the following theorem can be found in standard textbooks on the FEM, such as [30].

**Theorem 2.1** [1] Let  $u \in C^2(T)$ . Then we have the following interpolation error estimates:

$$\|u - \pi u\|_{L^2(T)} \leq C_1 h_T^2 \|D^2u\|_{L^2(T)}, \quad \|D(u - \pi u)\|_{L^2(T)} \leq C_2 h_T \|D^2u\|_{L^2(T)},$$

where  $C_1$  and  $C_2$  are positive constants that do not depend on the mesh size  $h_T$  or the function  $u$ , but depends on the ratio of the smallest to the largest interior angles of the triangle  $T$ .

Next, we provide the error estimates for the continuous piecewise interpolant defined by  $\pi u = \sum_{i=1}^N u(N_i) \phi_i$ .

**Theorem 2.2** Let  $u \in C^2(\Omega)$ . Then the global interpolation error satisfies the following estimates

$$\|u - \pi u\|_{L^2(\Omega)}^2 \leq C_1 \sum_{T \in \mathcal{T}_h} h_T^4 \|D^2u\|_{L^2(T)}^2, \quad \|D(u - \pi u)\|_{L^2(\Omega)}^2 \leq C_2 \sum_{T \in \mathcal{T}_h} h_T^2 \|D^2u\|_{L^2(T)}^2,$$

where  $C_1$  and  $C_2$  are positive constants (independent of the solution  $u$  and the mesh size  $h_T$ ), but depend on the ratio of the smallest to the largest interior angles of the triangles in the mesh  $\mathcal{T}_h$ . Here, the global  $H^1$ -seminorm is defined as  $\|u - \pi u\|_{H^1(\Omega)}^2 = \|D(u - \pi u)\|_{L^2(\Omega)}^2 = \sum_{T \in \mathcal{T}_h} \|D(u - \pi u)\|_{L^2(T)}^2$ . Next, we introduce the  $L^2$ -projection, which is fundamental to the theoretical error analysis of the FEM.

## 2.4 $L^2$ -projection

Suppose that  $\Omega \subset \mathbb{R}^2$ . Then, for  $u \in L^2(\Omega) = \left\{ v : \int_{\Omega} v^2(x, y) dx dy < \infty \right\}$ , the

$L^2$ -projection  $P_h : L^2(\Omega) \rightarrow V_h$  with  $V_h = \{v \in C(\Omega) : v|_T \in P^1(T), \forall T \in \mathcal{T}_h\}$  is defined by finding  $P_h u \in V_h$  such that  $P_h u \in V_h$  such that [1]

$$\int_{\Omega} (u - P_h u) v \, dx dy = 0, \quad \forall v \in V_h.$$

With the basis  $\{\phi_i\}_{i=1}^N$  chosen for  $V_h$ , the condition can be expressed as the following linear system:

$$\int_{\Omega} (u - P_h u) \phi_i \, dx dy = 0, \quad i = 1, 2, \dots, N.$$

Since  $P_h u \in V_h$ , it can be expressed as  $P_h u = \sum_{j=1}^N c_j \phi_j(x, y)$ , where the coefficients  $c_j \in \mathbb{R}$ . The coefficients  $c_j, j = 1, 2, 3$  satisfy

$$\sum_{j=1}^N c_j \int_{\Omega} \phi_i \phi_j \, dx dy = \int_{\Omega} u \phi_i \, dx dy, \quad i = 1, 2, \dots, N,$$

and can equivalently be written in matrix form as:

$$M \mathbf{c} = \mathbf{b}, \quad \text{where } \mathbf{c} = [c_1, c_2, \dots, c_N]^T,$$

and the entries of the  $N \times N$  matrix  $M$  and the  $N \times 1$  vector  $\mathbf{b}$  are determined by

$$m_{ij} = \int_{\Omega} \phi_i \phi_j \, dx dy, \quad b_i = \int_{\Omega} u \phi_i \, dx dy, \quad i, j = 1, 2, \dots, N.$$

In practice, integrals are usually approximated using a quadrature rule. More generally, a function  $f$  defined on a triangle  $T$  can be approximated as follows:

$$\int_T f(x, y) \, dx dy \approx \sum_{i=1}^N w_i f(\bar{N}_i),$$

where  $w_i$  are the quadrature weights and  $\bar{N}_i$  denote the associated quadrature points.

**Lemma 2.1** [1] The matrix  $M$  with entries  $m_{ij} = \int_{\Omega} \phi_i \phi_j \, dx dy$  is SPD (symmetric and positive definite).

**Theorem 2.3** [1] For all  $u \in L^2(\Omega)$ , the  $L^2$ -projection  $P_h u$  of  $u$  exists and is uniquely defined.

## 2.5 A priori error estimate

We now discuss the error estimates for the  $L^2$ -projection, which play a central role in the theoretical analysis of finite element approximations. The following result demonstrates that the  $L^2$ -projection gives the *best approximation* of  $u$  in the  $L^2$ -norm among all functions in the finite-dimensional space  $V_h$ .

**Theorem 2.4** [1] Suppose that  $u \in L^2(\Omega)$ . Let  $P_h u$  be the  $L^2$ -projection of  $u$  onto  $V_h$ . Then, the projection error satisfies

$$\|u - P_h u\|_{L^2(\Omega)} \leq \|u - v\|_{L^2(\Omega)}, \quad \forall v \in V_h.$$

The next a priori error estimate provides a quantitative bound on the projection error using the mesh size  $h_T$  and the smoothness of  $u$ . In practice, it implies that refining the mesh (reducing  $h_T$ ) improves the accuracy of the approximation.

**Theorem 2.5** [1] Suppose  $u \in C^2(\Omega)$ . Then, a positive constant  $C$ , independent of the mesh parameter, exists such that

$$\|u - P_h u\|_{L^2(\Omega)}^2 \leq C \sum_{T \in \mathcal{T}_h} h_T^4 \|D^2 u\|_{L^2(T)}^2.$$

## 2.6 The FEM for elliptic problems

The FEM was initially developed to approximate solutions to complex problems in elasticity and structural mechanics, often modeled by elliptic PDEs on geometrically intricate domains. Since its inception, the method has been extended to a broad range of additional applications.

To simplify the chapter, we focus on the two-dimensional elliptic model problem (1) and look for a (weak) solution,  $u$ , in the vector space

$$V = H^1(\Omega) = \left\{ v \in L^2(\Omega) : \|v\|_{L^2(\Omega)} < \infty, \|\nabla v\|_{L^2(\Omega)} < \infty \right\}.$$

To obtain the variational formulation, we multiply (1) by a smooth test function  $v \in V$ , integrate over the domain  $\Omega$ , and use Green's formula, yielding

$$\begin{aligned} \int_{\Omega} f v \, dx dy &= - \int_{\Omega} \nabla \cdot (a \nabla u) v \, dx dy + \int_{\Omega} b u v \, dx dy \\ &= \int_{\Omega} a \nabla u \cdot \nabla v \, dx dy - \int_{\partial \Omega} v \mathbf{n} \cdot (a \nabla u) \, ds + \int_{\Omega} b u v \, dx dy \\ &= \int_{\Omega} a \nabla u \cdot \nabla v \, dx dy + \int_{\Omega} b u v \, dx dy + \int_{\partial \Omega} (\epsilon(u - g) - q) v \, ds. \end{aligned}$$

Then, the problem in weak form reads: find  $u \in V$  satisfying [1]

$$\int_{\Omega} a \nabla u \cdot \nabla v \, dx dy + \int_{\Omega} b u v \, dx dy + \epsilon \int_{\partial \Omega} u v \, ds = \int_{\Omega} f v \, dx dy + \int_{\partial \Omega} (\epsilon g + q) v \, ds, \quad v \in V. \quad (2)$$

We formulate the FEM using the weak formulation (2). The two-dimensional FEM is then constructed as follows: determine  $u_h \in V_h$  such that (2). The FEM in 2D read as follows: find  $u_h \in V_h = \{v \in V : v|_T \in P^1(T), \forall T \in \mathcal{T}_h\}$  satisfying

$$\int_{\Omega} a \nabla u_h \cdot \nabla v \, dx dy + \int_{\Omega} b u_h v \, dx dy + \epsilon \int_{\partial \Omega} u_h v \, ds = \int_{\Omega} f v \, dx dy + \epsilon \int_{\partial \Omega} g v \, ds, \quad v \in V_h. \quad (3)$$

**Implementation:** Here, we discuss the practical implementation of the FEM. For simplicity, let  $a = 1$  and  $b = g = q = 0$ . Substituting the FEM solution expansion

$u_h = \sum_{j=1}^N c_j \phi_j$  into (3) and taking the test function  $v = \phi_i$ , we get

$$\sum_{j=1}^N c_j \left( \int_{\Omega} \nabla \phi_i \cdot \nabla \phi_j \, dx dy + \int_{\partial \Omega} \epsilon \phi_i \phi_j \, ds \right) = \int_{\Omega} f \phi_i \, dx dy, \quad i = 1, 2, \dots, N.$$

This results in a linear system of the form:

$$(A + R)\mathbf{c} = \mathbf{b},$$

where the unknown vector  $\mathbf{c} = [c_1, c_2, \dots, c_N]^T \in \mathbb{R}^N$  is to be determined. The entries of the stiffness matrix  $A \in \mathbb{R}^{N \times N}$ , the boundary matrix  $R \in \mathbb{R}^{N \times N}$ , and the load vector  $\mathbf{b} \in \mathbb{R}^N$  are defined by

$$a_{ij} = \int_{\Omega} \nabla \phi_j \cdot \nabla \phi_i \, dxdy, \quad r_{ij} = \epsilon \int_{\partial\Omega} \phi_j \phi_i \, ds, \quad b_i = \int_{\Omega} f \phi_i \, dxdy, \quad i, j = 1, 2, \dots, N.$$

**Assembly of the stiffness matrix  $A$ :** We begin by identifying the local contributions from each triangle  $T \in \mathcal{T}_h$ . For a single triangle  $T$  we define

$$a_{ij}^T = \int_T \nabla \phi_j \cdot \nabla \phi_i \, dxdy, \quad i, j = 1, 2, 3,$$

where  $N_i = (x_i, y_i)$  are the vertices of  $T$  and  $\phi_i$  are the associated hat functions satisfying  $\phi_j(N_i) = \delta_{ij}$ . Each linear hat function  $\phi_i$  can be written as  $\phi_i(x, y) = \alpha_i + \beta_i x + \gamma_i y$ ,  $i = 1, 2, 3$ , where the coefficients  $\alpha_i, \beta_i, \gamma_i$  are computed by solving  $B\alpha_i = \mathbf{e}_i$ ,  $i = 1, 2, 3$ , with

$$B = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix}, \quad \alpha_i = \begin{bmatrix} \alpha_i \\ \beta_i \\ \gamma_i \end{bmatrix}, \quad \mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Since  $\nabla \phi_i = [\beta_i, \gamma_i]^T$ , the local stiffness entries are

$$a_{ij}^T = \int_T (\beta_i \beta_j + \gamma_i \gamma_j) \, dxdy = (\beta_i \beta_j + \gamma_i \gamma_j) |T|, \quad i, j = 1, 2, 3,$$

with  $|T|$  being the area of the triangle  $T$ .

**Assembly of the boundary matrix  $R$ :** Let  $\Gamma_h^{\text{out}}$  be the collection of all boundary edges of the partition, i.e.,  $\Gamma_h^{\text{out}} = \{E : E = T \cap \partial\Omega, \forall T \in \mathcal{T}_h\}$ . Let  $\epsilon$  be a constant on each edge  $E \in \Gamma_h^{\text{out}}$ . For each such edge  $E$ , we define  $R^E \in \mathbb{R}^{2 \times 2}$  with entries

$$r_{ij}^E = \int_E \epsilon \phi_j \phi_i \, ds = \epsilon \frac{|E|}{6} (1 + \delta_{ij}), \quad i, j = 1, 2,$$

where  $\delta_{ij}$  is the Kronecker delta symbol and  $|E|$  denotes the length of the edge  $E$ .

**Assembly of the load vector  $\mathbf{b}$ :** To approximate the element-wise contributions we employ a corner quadrature rule. For each  $T \in \mathcal{T}_h$  we obtain

$$b_i^T = \int_T f \phi_i \, dxdy \approx \frac{|T|}{3} f(N_i), \quad i = 1, 2, 3.$$

Finally, once  $A$ ,  $R$ , and  $\mathbf{b}$  have been assembled, we solve the global system  $(A + R)\mathbf{c} = \mathbf{b}$  and reconstruct the finite element solution

$$\text{as } u_h(x, y) = \sum_{j=1}^N c_j \phi_j(x, y).$$

## 2.7 The Dirichlet problem

We now discuss the Dirichlet BVP: find  $u$  satisfying [7]

$$-\Delta u = f(x, y), \quad (x, y) \in \Omega, \quad u = g, \quad (x, y) \in \partial\Omega, \quad (4)$$

where the input functions  $f \in L^2(\Omega)$  and  $g \in C(\partial\Omega)$ . We would like to find a weak solution  $u$  in the affine space  $V_g = \{v \in V : v|_{\partial\Omega} = g\}$ . To derive the variational

formulation, we multiply (4) by a smooth test function  $v \in V_0 = \{v \in V : v|_{\partial\Omega} = 0\}$  and integrate over  $\Omega$ :

$$\int_{\Omega} f v \, dx dy = - \int_{\Omega} v \Delta u \, dx dy = \int_{\Omega} \nabla u \cdot \nabla v \, dx dy - \int_{\partial\Omega} v \nabla u \cdot \mathbf{n} \, ds = \int_{\Omega} \nabla u \cdot \nabla v \, dx dy,$$

since  $v|_{\partial\Omega} = 0$ . Thus, the variational formulation consists of finding  $u \in V_g$  with [7]

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx dy = \int_{\Omega} f v \, dx dy, \quad \text{for all } v \in V_0.$$

**Finite element formulation:** Suppose that  $g$  is piecewise linear on the boundary  $\partial\Omega$ . Then the FEM consists of finding  $u_h \in V_{h,g} = \{v \in V_h : v|_{\partial\Omega} = g\}$  satisfying

$$\int_{\Omega} \nabla u_h \cdot \nabla v \, dx dy = \int_{\Omega} f v \, dx dy, \quad \forall v \in V_{h,0} = \{v \in V_h : v|_{\partial\Omega} = 0\}.$$

**Matrix formulation:** If  $N$  is the total number of nodes and  $J$  is the total the number of boundary nodes. Then the discrete system can be expressed in the following block-matrix form [1, 7]:

$$\begin{bmatrix} A_{0,0} & A_{0,g} \\ A_{g,0} & A_{g,g} \end{bmatrix} \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_0 \\ \mathbf{b}_1 \end{bmatrix},$$

where  $A_{0,0}$  is  $(N - J) \times (N - J)$  matrix,  $A_{0,g}$  is  $(N - J) \times J$  matrix,  $A_{g,0}$  is  $J \times (N - J)$  matrix, and  $A_{g,g}$  is  $J \times J$  matrix. Observe that the vector  $\mathbf{c}_1 \in \mathbb{R}^J$  is prescribed since it has the prescribed boundary values of  $g$  at the boundary vertices. Consequently, the reduced system to be solved is  $A_{0,0}\mathbf{c}_0 = \mathbf{b}_0 - A_{0,g}\mathbf{c}_1$ , where the unknown vector  $\mathbf{c}_0 \in \mathbb{R}^{N-J}$  correspond to the interior nodes.

## 2.8 Error estimates

Let us examine the following model elliptic BVP. The task is to find a function  $u = u(x,y)$  satisfying

$$-\Delta u = f, \quad \text{in } \Omega, \quad u = 0, \quad \text{on } \partial\Omega.$$

The variational formulation consists of finding  $u \in V_0 = \{v \in V : v|_{\partial\Omega} = 0\}$  satisfying

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx dy = \int_{\Omega} f v \, dx dy, \quad \text{for all } v \in V_0.$$

Let  $V_h = \{v \in V : v|_T \in P^1(T), \forall T \in \mathcal{T}_h\}$  and  $V_{h,0} = V_h \cap V_0$ . The FEM consists of finding  $u_h \in V_{h,0}$  satisfying

$$\int_{\Omega} \nabla u_h \cdot \nabla v \, dx dy = \int_{\Omega} f v \, dx dy, \quad \forall v \in V_{h,0}.$$

Expanding  $u_h$  in the nodal basis  $\{\phi_j\}_{j=1}^N$  as  $u_h = \sum_{j=1}^N c_j \phi_j$ , and testing with  $v = \phi_i$  gives

$$\sum_{j=1}^N c_j \left( \int_{\Omega} \nabla \phi_i \cdot \nabla \phi_j \, dx dy \right) = \int_{\Omega} f \phi_i \, dx dy, \quad i = 1, 2, \dots, N.$$

This leads to a linear system of equations, which can be expressed as  $A\mathbf{c} = \mathbf{b}$ , where  $A$  is the  $N \times N$  stiffness matrix and  $\mathbf{b}$  is the  $N \times 1$  load vector. The entries of  $A$  and  $\mathbf{b}$  are given by

$$a_{ij} = \int_{\Omega} \nabla \phi_i \cdot \nabla \phi_j \, dxdy, \quad b_i = \int_{\Omega} f \phi_i \, dxdy, \text{ for } i, j = 1, 2, \dots, N.$$

**Theorem 2.6** [1, 7] The stiffness matrix  $A$ , whose entries are given by  $a_{ij} = \int_{\Omega} \nabla \phi_j \cdot \nabla \phi_i \, dxdy$  is symmetric and positive definite.

**Theorem 2.7 (Galerkin orthogonality condition** [1, 7],) Suppose that  $u \in V_0$  is the weak solution of the weak form and  $u_h \in V_{h,0}$  its finite element approximation. Then

$$\int_{\Omega} \nabla(u - u_h) \cdot \nabla v \, dxdy = 0, \quad \forall v \in V_{h,0}.$$

For convenience, we introduce the *energy norm* on the space  $V_0$  by

$$|||v|||^2 = \int_{\Omega} |\nabla v|^2 \, dxdy = \int_{\Omega} \nabla v \cdot \nabla v \, dxdy = ||\nabla v||_{L^2(\Omega)}^2.$$

Error estimates can be classified into two main types:

*A priori error estimates:* estimates that bound the error in terms of norms of the exact solution  $u$ ;

*A posteriori error estimates:* estimates that bound the error using quantities derived solely from the computed finite element solution  $u_h$ .

**Theorem 2.8 (A priori error estimate** [1, 7]) Assume that  $u \in V_0$  is the weak solution and  $u_h \in V_{h,0}$  denoting its finite element approximation. Then

$$|||u - u_h||| \leq |||u - v|||, \quad \forall v \in V_{h,0}.$$

**Theorem 2.9 (Interpolation error estimate** [1, 7],) Suppose that  $u \in V_0$  is the weak solution of the weak form and  $u_h \in V_{h,0}$  is its finite element approximation. Let  $u \in C^2(\Omega)$ . Then there exists a positive constant  $C$  independent of  $u$  and  $h_T$  such that [1]

$$||u - u_h||_{L^2(\Omega)}^2 \leq C \sum_{T \in \mathcal{T}_h} h_T^4 ||D^2 u||_{L^2(T)}^2.$$

## 2.9 The Neumann problem

Here, we study the Neumann boundary-value problem, which consists of finding  $u = u(x, y)$  satisfying [7]

$$-\Delta u = f, \quad (x, y) \in \Omega \quad \text{subject to} \quad \mathbf{n} \cdot \nabla u = g, \quad (x, y) \in \partial\Omega, \quad (5)$$

where we assume the functions  $f \in L^2(\Omega)$  and  $g \in C(\partial\Omega)$ . Our goal is to find a weak solution in the vector space  $V = \{v : ||v||_{L^2(\Omega)} < \infty, ||\nabla v||_{L^2(\Omega)} < \infty\}$ . To derive the weak form, we multiply (5) by a smooth test function  $v \in V$ , integrate over  $\Omega$ , and apply Green's identity to obtain

$$\begin{aligned}\int_{\Omega} f v \, dx dy &= - \int_{\Omega} v \, \Delta u \, dx dy = \int_{\Omega} \nabla u \cdot \nabla v \, dx dy - \int_{\partial \Omega} \mathbf{n} \cdot \nabla u \, v \, ds \\ &= \int_{\Omega} \nabla u \cdot \nabla v \, dx dy - \int_{\partial \Omega} v g \, ds.\end{aligned}$$

Thus, the weak formulation reads as follows: determine  $u$  in  $V$  satisfying

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx dy = \int_{\Omega} f v \, dx dy + \int_{\partial \Omega} v g \, ds, \quad \forall v \in V.$$

**Solvability and compatibility condition:** To ensure solvability, consider  $v = 1$ :

$$0 = \int_{\Omega} \nabla u \cdot \nabla 1 \, dx dy = \int_{\Omega} f \, dx dy + \int_{\partial \Omega} g \, ds.$$

Consequently, the problem requires the satisfaction of the following *compatibility condition*:

$$\int_{\Omega} f \, dx dy + \int_{\partial \Omega} g \, ds = 0.$$

**Non-uniqueness and normalization** [1, 7]: If  $u$  is a solution, then  $u + c$  is also a solution for any constant  $c \in \mathbb{R}$ . Thus, the solution is only defined up to an additive constant. A common way to get uniqueness is to impose the normalization

$\int_{\Omega} u \, dx dy = 0$ . Accordingly, we define the new subspace

$$\hat{V} = \left\{ v \in V : \int_{\Omega} v \, dx dy = 0 \right\},$$

which consists of all functions that have zero mean value. This new space, known as a quotient space, ensures the uniqueness of the weak solution. The Neumann problem can now be stated as follows: determine  $u \in \hat{V}$  such that

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx dy = \int_{\Omega} f v \, dx dy + \int_{\partial \Omega} v g \, ds, \quad \forall v \in \hat{V}.$$

Define  $\hat{V}_h \subset \hat{V}$  to be the space of continuous piecewise linear polynomials with zero mean. The finite element method is defined as follows: determine  $u_h \in \hat{V}_h$  satisfying

$$\int_{\Omega} \nabla u_h \cdot \nabla v \, dx dy = \int_{\Omega} f v \, dx dy + \int_{\partial \Omega} v g \, ds, \quad \forall v \in \hat{V}_h.$$

## 2.10 FEM for mixed (Dirichlet-Neumann) boundary conditions

Here, we explain how mixed boundary conditions are treated within the 2D finite element framework. Let  $\Omega$  be the computational domain in either  $\mathbb{R}^2$  or  $\mathbb{R}^3$ , and suppose its boundary is partitioned into two disjoint parts,  $\partial \Omega = \Gamma_1 \cup \Gamma_2$  with  $\Gamma_1 \cap \Gamma_2 = \emptyset$ . We consider the following BVP:

$$-\nabla \cdot (a(\mathbf{x})\nabla u) = f(\mathbf{x}), \text{ in } \mathbf{x} \in \Omega, \text{ subject to } u = 0, \text{ for } \mathbf{x} \in \Gamma_1, \text{ and } \mathbf{n} \cdot \nabla u = 0, \text{ for } \mathbf{x} \in \Gamma_2, \quad (6)$$

where  $a(\mathbf{x})$  is a given coefficient function and  $f \in L^2(\Omega)$ . Dirichlet conditions are referred to as *essential boundary conditions* since they must be imposed explicitly in the finite element formulation. In contrast, Neumann conditions are called *natural boundary conditions*, which are automatically incorporated into the weak formulation. Accordingly, we introduce the following vector space of test functions

$$\hat{V} = \{v \in C^2(\overline{\Omega}) : v(\mathbf{x}) = 0 \text{ for all } \mathbf{x} \in \Gamma_1\}.$$

We multiply (6) by a smooth test function  $v \in \hat{V}$ , integrate over the domain  $\Omega$ , and use Green's Theorem to get

$$\begin{aligned} \int_{\Omega} f v \, dx dy &= - \int_{\Omega} \nabla \cdot (a(\mathbf{x})\nabla u) v \, dx dy \\ &= \int_{\Omega} a(\mathbf{x}) \nabla u \cdot \nabla v \, dx dy - \int_{\partial\Omega} a(\mathbf{x}) \mathbf{n} \cdot \nabla u \, v \, ds \\ &= \int_{\Omega} a(\mathbf{x}) \nabla u \cdot \nabla v \, dx dy - \int_{\Gamma_1} a(\mathbf{x}) \mathbf{n} \cdot \nabla u \, v \, ds - \int_{\Gamma_2} a(\mathbf{x}) \mathbf{n} \cdot \nabla u \, v \, ds. \end{aligned}$$

Because  $v = 0$  on  $\Gamma_1$  and also  $\mathbf{n} \cdot \nabla u = 0$  on  $\Gamma_2$ , the boundary terms vanish. Therefore, the weak formulation of (6) is: find  $u \in \hat{V}$  satisfying [1]

$$\int_{\Omega} a(\mathbf{x}) \nabla u \cdot \nabla v \, dx dy = \int_{\Omega} f v \, dx dy, \quad \forall v \in \hat{V}. \quad (7)$$

We now focus on two-dimensional polygonal domains. To implement the FEM, we select a discrete subspace of  $\hat{V}$ . Due to the presence of mixed boundary conditions, certain points on  $\partial\Omega$  mark transitions between Dirichlet and Neumann boundaries. We assume that the mesh is constructed so that these transition points coincide with mesh nodes, which are then assigned to  $\Gamma_1$ . This ensures that  $\Gamma_1$  includes its endpoints. Accordingly, the finite element trial space is defined as

$$V_h = \{v \in C(\overline{\Omega}) : v \in P^1(T), \forall T \in \mathcal{T}_h, v(N) = 0 \text{ for all nodes } N \in \Gamma_1\}.$$

A convenient basis for the finite-dimensional space  $V_h$  is given by the standard hat functions associated with all interior nodes and boundary nodes not on  $\Gamma_1$ . It is important to note that if the problem involves only Neumann boundary conditions, the resulting stiffness matrix becomes singular. This occurs because the PDE either has no solution or admits infinitely many solutions differing by an additive constant. Therefore, special care is required to ensure uniqueness and to obtain a meaningful solution for the resulting discrete system.

## 2.11 Shifting the data method

In two dimensions, nonhomogeneous Dirichlet boundary conditions are handled using the shifting-the-data method, which employs a suitably chosen auxiliary function. By contrast, nonhomogeneous Neumann conditions are incorporated

directly into the weak formulation. In both cases, the primary effect is a modification of the load vector in the finite element method.

### 2.11.1 Inhomogeneous Dirichlet conditions on a rectangle

The key idea is to transform a problem with inhomogeneous Dirichlet conditions into an equivalent problem with homogeneous conditions. This approach is analogous to the one-dimensional case, although, in two dimensions, constructing a function that satisfies the prescribed boundary data is more involved.

To illustrate the method, we focus on the following inhomogeneous boundary-value problem [1]:

$$-\Delta u = f(\mathbf{x}), \text{ in } \Omega = (0, a) \times (0, b), \text{ subject to } u(\mathbf{x}) = g(\mathbf{x}) = \begin{cases} g_1(\mathbf{x}), & \mathbf{x} \in \Gamma_1, \\ g_2(\mathbf{x}), & \mathbf{x} \in \Gamma_2, \\ g_3(\mathbf{x}), & \mathbf{x} \in \Gamma_3, \\ g_4(\mathbf{x}), & \mathbf{x} \in \Gamma_4, \end{cases} \quad (8)$$

where  $\Gamma_1, \Gamma_2, \Gamma_3$ , and  $\Gamma_4$  are used to denote the bottom, right, top, and left edges of the rectangle  $\Omega = (0, a) \times (0, b)$ , respectively. We assume throughout that the boundary conditions are consistent *i.e.*,

$$g_1(0) = g_4(0), \quad g_1(a) = g_2(0), \quad g_2(b) = g_3(a), \quad g_3(0) = g_4(b).$$

If it is possible to construct a function  $w = w(\mathbf{x})$ , defined on  $\overline{\Omega}$ , such that  $w(\mathbf{x}) = g(\mathbf{x})$ , for all  $\mathbf{x} \in \partial\Omega$ , then the new function  $v = u - w$  satisfies

$$-\Delta v = -\Delta u + \Delta w = f(\mathbf{x}) + \Delta w = \hat{f}(\mathbf{x}),$$

with homogeneous boundary conditions  $v(\mathbf{x}) = 0$  on  $\partial\Omega$ . Thus, instead of solving (8), we solve the equivalent homogeneous problem

$$-\Delta v = \hat{f}(\mathbf{x}), \quad \text{in } \Omega, \quad v(\mathbf{x}) = 0, \quad \text{on } \partial\Omega, \quad (9)$$

and recover the original solution via  $u = v + w$ .

We now turn to the construction of a suitable function,  $w$ , that satisfies the given boundary conditions. First, observe that there exists a bilinear polynomial of the form  $q(x, y) = c_0 + c_1x + c_2y + c_3xy$  that interpolates the given boundary data at the four corners of the domain  $\Omega$ :

$$\begin{aligned} g_1(0) = q(0, 0) = g_4(0), \quad g_1(a) = q(a, 0) = g_2(0), \quad g_2(b) = q(a, b) = g_3(a), \\ g_3(0) = q(0, b) = g_4(b). \end{aligned}$$

A straightforward calculation gives

$$\begin{aligned} c_0 = g_1(0), \quad c_1 = \frac{g_1(a) - g_1(0)}{a}, \quad c_2 = \frac{g_4(b) - g_4(0)}{b}, \\ c_3 = \frac{g_2(b) + g_1(0) - g_1(a) - g_4(b)}{ab}. \end{aligned}$$

Next, we modify the boundary data by subtracting appropriate linear functions so that the corrected functions vanish at the endpoints:

$$h(\mathbf{x}) = \begin{cases} h_1(x) = g_1(x) - (g_1(0) + \frac{g_1(a) - g_1(0)}{a}x), & \mathbf{x} \in \Gamma_1, \\ h_2(y) = g_2(y) - (g_2(0) + \frac{g_2(b) - g_2(0)}{b}y), & \mathbf{x} \in \Gamma_2, \\ h_3(x) = g_3(x) - (g_3(0) + \frac{g_3(a) - g_3(0)}{a}x), & \mathbf{x} \in \Gamma_3, \\ h_4(y) = g_4(y) - (g_4(0) + \frac{g_4(b) - g_4(0)}{b}y), & \mathbf{x} \in \Gamma_4. \end{cases}$$

By construction, each  $h_i$  vanishes at the two endpoints of its edge:

$$h_1(0) = h_1(a) = h_2(0) = h_2(b) = h_3(0) = h_3(a) = h_4(0) = h_4(b) = 0.$$

Finally, we define

$$w(x, y) = (c_0 + c_1x + c_2y + c_3xy) + \left( h_4(y) + \frac{h_2(y) - h_4(y)}{a}x \right) + \left( h_1(x) + \frac{h_3(x) - h_1(x)}{b}y \right).$$

The second term provides interpolation between the boundary data on  $\Gamma_1$  and  $\Gamma_3$ , while the third term interpolates between the edges  $\Gamma_2$  and  $\Gamma_4$ . To prevent overlap, the modified data  $h_i$  are chosen to vanish at the corners, and the first term restores the correct corner values. It can be readily verified that the resulting function  $w$  satisfies the prescribed inhomogeneous Dirichlet conditions.

### 2.11.2 Inhomogeneous Neumann boundary conditions on a rectangle

The data-shifting technique can also be used to transform a boundary-value problem with inhomogeneous Neumann conditions into one with homogeneous Neumann conditions. Compared to the Dirichlet case, this construction requires additional care. We examine the boundary-value problem given by [1]:

$$-\Delta u = f(\mathbf{x}), \quad \text{in } \Omega = (0, a) \times (0, b), \quad \text{subject to } \mathbf{n} \cdot \nabla u = g(\mathbf{x}) = \begin{cases} g_1(x), & \mathbf{x} \in \Gamma_1, \\ g_2(y), & \mathbf{x} \in \Gamma_2, \\ g_3(x), & \mathbf{x} \in \Gamma_3, \\ g_4(y), & \mathbf{x} \in \Gamma_4, \end{cases} \quad (10)$$

where, as before,  $\Gamma_1, \Gamma_2, \Gamma_3$ , and  $\Gamma_4$  are the bottom, right, top, and left edges of the rectangular domain  $\Omega = (0, a) \times (0, b)$ , respectively. Explicitly, the Neumann conditions are

$$-u_x(0, y) = g_4(y), \quad u_x(a, y) = g_2(y), \quad -u_y(x, 0) = g_1(x), \quad u_y(x, b) = g_3(x).$$

A consistency condition arises from the equality of mixed partial derivatives  $u_{xy} = u_{yx}$ . For example,

$$-u_{xy}(x, 0) = g_1'(x), \quad -u_{yx}(0, y) = g_4'(y),$$

which yields  $g_1'(0) = g_4'(0)$ . Repeating this argument at the other corners leads to the set of compatibility conditions

$$g_1'(0) = g_4'(0), \quad -g_3'(0) = g_4'(b), \quad -g_1'(a) = g_2'(0), \quad g_2'(b) = g_3'(a). \quad (11)$$

We assume throughout that these conditions are satisfied. The main idea is again to design a suitable function,  $w$ , that satisfies the given Neumann conditions. As in the Dirichlet case, we will interpolate boundary data in such a way that contributions from different edges do not interfere with each other. A useful ingredient is the one-dimensional function

$$\psi(x) = -\alpha x + \frac{\alpha + \beta}{2a} x^2, \quad (12)$$

which satisfies  $\psi'(0) = -\alpha$  and  $\psi'(a) = \beta$ . As a first step, the boundary data are modified to ensure that their derivatives vanish at the endpoints. Specifically, each  $g_i$  is transformed into  $h_i$ , with  $h_1'(0) = 0 = h_1'(a)$ , and also for  $h_2$ ,  $h_3$ , and  $h_4$ . Since these endpoint derivatives correspond to mixed partial derivatives of  $u$ , we then seek a function  $q(x, y)$  such that

$$u_{xy}(0, 0) = -g_1'(0), \quad u_{xy}(a, b) = -g_2'(b), \quad u_{xy}(a, 0) = -g_1'(a), \quad u_{xy}(0, b) = -g_3'(0).$$

This can be achieved with  $q(x, y) = c_0xy + c_1x^2y + c_2xy^2 + c_3x^2y^2$ , where the coefficients are determined by the corner conditions:

$$\begin{aligned} c_0 &= -g_1'(0), c_1 = \frac{1}{2a}(g_1'(0) - g_1'(a)), c_2 = \frac{1}{2b}(g_1'(0) + g_3'(0)), \\ c_3 &= \frac{1}{4ab}(g_2'(b) + g_1'(a) - g_1'(0) - g_3'(0)). \end{aligned}$$

Next, we define the modified data along each boundary edge:

$$h_1(x) = g_1(x) + c_0x + c_1x^2, \quad h_2(y) = g_2(y) - (c_0 + 2ac_1)y - (c_2 + 2ac_3)y^2,$$

$$h_3(x) = g_3(x) - (c_0 + 2bc_2)x - (c_1 + 2bc_3)x^2, \quad h_4(y) = g_4(y) + c_0y + c_2y^2.$$

Finally, we interpolate to construct  $w$ :

$$w(x, y) = q(x, y) - h_1(x)y + \frac{h_3(x) + h_1(x)}{2b}y^2 - h_4(y)x + \frac{h_2(y) + h_4(y)}{2a}x^2y.$$

By direct verification,  $w$  satisfies the original Neumann boundary conditions.

## 2.12 Eigenvalue problem

We now turn our attention to an eigenvalue problem for the Laplacian subject to homogeneous Neumann boundary conditions. Specifically, we want to find  $\lambda \in \mathbb{R}$  and a nontrivial function  $u$  satisfying

$$-\Delta u = \lambda u, \quad \text{in } \Omega, \quad \mathbf{n} \cdot \nabla u = 0, \quad \text{on } \partial\Omega. \quad (13)$$

To derive the weak formulation, we multiply (13) by a smooth test function  $v$  in  $V$ , we integrate over the domain  $\Omega$ , and we invoke Green's Theorem:

$$\lambda \int_{\Omega} uv \, dx dy = - \int_{\Omega} \Delta u v \, dx dy = \int_{\Omega} \nabla u \cdot \nabla v \, dx dy - \int_{\partial\Omega} \mathbf{n} \cdot \nabla u v \, ds.$$

Since  $\mathbf{n} \cdot \nabla u = 0$  on  $\partial\Omega$ , the boundary term vanishes, leaving

$$\lambda \int_{\Omega} uv \, dxdy = \int_{\Omega} \nabla u \cdot \nabla v \, dxdy.$$

The variational formulation reads: determine  $u \in V$ ,  $u \neq 0$ , and  $\lambda \in \mathbb{R}$  such that

$$\int_{\Omega} \nabla u \cdot \nabla v \, dxdy = \lambda \int_{\Omega} uv \, dxdy, \quad \forall v \in V. \quad (14)$$

Let  $V_h = \{v \in V : v|_T \in P^1(T), \forall T \in \mathcal{T}_h\}$  be the classical piecewise linear finite element space. The discrete problem becomes: Find  $\lambda_h \in \mathbb{R}$  and  $u_h \in V_h$  with  $u_h \neq 0$  such that

$$\int_{\Omega} \nabla u_h \cdot \nabla v \, dxdy = \lambda_h \int_{\Omega} u_h v \, dxdy, \quad \forall v \in V_h. \quad (15)$$

Expanding the approximate solution as  $u_h = \sum_{j=1}^N c_j \phi_j$  and testing with  $v = \phi_i$  yields

$$\sum_{j=1}^N c_j \int_{\Omega} \nabla \phi_j \cdot \nabla \phi_i \, dxdy = \lambda_h \sum_{j=1}^N c_j \int_{\Omega} \phi_i \phi_j \, dxdy, \quad i = 1, 2, \dots, N,$$

which can be written as a classical eigenvalue problem:

$$A\mathbf{c} = \lambda_h M\mathbf{c},$$

where the entries of the mass matrix  $M$  are  $m_{ij} = \int_{\Omega} \phi_i \phi_j \, dxdy$ , the entries of the stiffness matrix  $A$  are  $a_{ij} = \int_{\Omega} \nabla \phi_j \cdot \nabla \phi_i \, dxdy$ , and  $\mathbf{c} = [c_1, c_2, \dots, c_N]^T$ .

### 2.13 The FEM for convection-diffusion problems

Here, we examine the following reaction-convection-diffusion problem, where the goal is to determine a function  $u$  such that [1]

$$-\nabla \cdot (a \nabla u) + \mathbf{b} \cdot \nabla u + cu = f, \quad \mathbf{x} \in \Omega, \text{ subject to } u = 0, \quad \mathbf{x} \in \partial\Omega, \quad (16)$$

where  $a > 0$  is the diffusion coefficient,  $\mathbf{b}$  is a given convection field, and  $c \geq 0$  is a reaction coefficient. We determine a weak solution  $u \in V_0 = \{v \in V : v|_{\partial\Omega} = 0\}$ . Multiplying (16) by  $v$  in  $V_0$ , integrating over the domain  $\Omega$ , and applying Green's Theorem to the diffusion term, we obtain

$$\begin{aligned} \int_{\Omega} fv \, dxdy &= - \int_{\Omega} v \nabla \cdot (a \nabla u) \, dxdy + \int_{\Omega} v \mathbf{b} \cdot \nabla u \, dxdy + \int_{\Omega} cuv \, dxdy \\ &= \int_{\Omega} a \nabla u \cdot \nabla v \, dxdy - \int_{\partial\Omega} \mathbf{n} \cdot (a \nabla u) v \, ds + \int_{\Omega} v \mathbf{b} \cdot \nabla u \, dxdy + \int_{\Omega} cuv \, dxdy \\ &= \int_{\Omega} a \nabla u \cdot \nabla v \, dxdy + \int_{\Omega} v \mathbf{b} \cdot \nabla u \, dxdy + \int_{\Omega} cuv \, dxdy, \end{aligned}$$

since  $v = 0$  on  $\partial\Omega$ . Observe that it is not necessary to integrate the convection term by parts. Consequently, the weak formulation consists of determining  $u$  in  $V_0$  with

$$\int_{\Omega} a \nabla u \cdot \nabla v \, dxdy + \int_{\Omega} v \mathbf{b} \cdot \nabla u \, dxdy + \int_{\Omega} cuv \, dxdy = \int_{\Omega} fv \, dxdy, \quad \forall v \in V_0.$$

Let  $V_h = \{v \in V : v|_T \in P^1(T), \forall T \in \mathcal{T}_h\}$  and  $V_{h,0} = V_h \cap V_0$ . Then we define the discrete problem as follows: Find  $u_h \in V_{h,0}$  such that

$$\int_{\Omega} a \nabla u_h \cdot \nabla v \, dx dy + \int_{\Omega} v \mathbf{b} \cdot \nabla u_h \, dx dy + \int_{\Omega} c u_h v \, dx dy = \int_{\Omega} f v \, dx dy, \quad \forall v \in V_{h,0}. \quad (17)$$

Expanding  $u_h$  in the nodal basis  $\{\phi_j\}_{j=1}^N$ ,  $u_h = \sum_{j=1}^N c_j \phi_j$ , and testing with  $v = \phi_i$  yields

$$\sum_{j=1}^N c_j \left( \int_{\Omega} a \nabla \phi_i \cdot \nabla \phi_j \, dx dy + \int_{\Omega} \phi_i \mathbf{b} \cdot \nabla \phi_j \, dx dy + \int_{\Omega} c \phi_i \phi_j \, dx dy \right) = \int_{\Omega} f \phi_i \, dx dy, \\ i = 1, 2, \dots, N.$$

Consequently, the problem is reduced to a linear system of the form  $(A + B + C)\mathbf{c} = \mathbf{d}$ , where the unknown vector  $\mathbf{c} = [c_1, c_2, \dots, c_N]^T \in \mathbb{R}^N$ , and the matrices  $A, B, C \in \mathbb{R}^{N \times N}$  together with the load vector  $\mathbf{d} \in \mathbb{R}^N$  are defined as [1, 7]

$$a_{ij} = \int_{\Omega} a \nabla \phi_i \cdot \nabla \phi_j \, dx dy, \quad b_{ij} = \int_{\Omega} \phi_i \mathbf{b} \cdot \nabla \phi_j \, dx dy, \quad c_{ij} = \int_{\Omega} c \phi_i \phi_j \, dx dy, \\ d_i = \int_{\Omega} f \phi_i \, dx dy,$$

for all  $i, j = 1, 2, \dots, N$ . It is important to note that the convection matrix  $B$  is generally nonsymmetric, i.e.  $b_{ji} \neq b_{ij}$ .

### 3. The FEM for the heat (or diffusion) equation

We design a FEM for the classical diffusion (heat) problem: determine  $u(\mathbf{x}, t)$  satisfying [1]

$$\frac{\partial u}{\partial t} - \Delta u = f, \quad \text{in } \Omega \subset \mathbb{R}^2, \quad t \in [0, T], \quad (18)$$

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \quad (19)$$

$$u(\mathbf{x}, t) = 0, \quad \text{on } \partial\Omega, \quad t \in [0, T]. \quad (20)$$

Our goal is to determine a weak solution  $u \in V_0 = \{v \in V : \|v\| < \infty, \|\nabla v\| < \infty, v|_{\partial\Omega} = 0\}$ . Multiplying (18) by a smooth test function  $v \in V_0$ , integrating over the domain  $\Omega$ , and applying Green's formula to the diffusion term, we get, for  $t \in [0, T]$ :

$$\int_{\Omega} f v \, d\mathbf{x} = \int_{\Omega} v \frac{\partial u}{\partial t} \, d\mathbf{x} + \int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x} - \int_{\partial\Omega} v \mathbf{n} \cdot \nabla u \, ds = \int_{\Omega} v \frac{\partial u}{\partial t} \, d\mathbf{x} + \int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x},$$

since  $v = 0$  on the boundary  $\partial\Omega$ . Hence, the variational formulation consists of finding  $u(\cdot, t) \in V_0$  such that [1]

$$\int_{\Omega} v \frac{\partial u}{\partial t} \, d\mathbf{x} + \int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x}, \quad \forall v \in V_0, \quad \forall t > 0. \quad (21)$$

Define  $V_h = \{v \in V : v|_T \in P^1(T), \forall T \in \mathcal{T}_h\}$  and  $V_{h,0} = V_h \cap V_0$ . The semi-discrete problem is defined as follows: determine  $u_h(\cdot, t) \in V_{h,0}$  satisfying

$$\int_{\Omega} \frac{\partial u_h}{\partial t} v \, d\mathbf{x} + \int_{\Omega} \nabla u_h \cdot \nabla v \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x}, \quad \forall v \in V_{h,0}. \quad (22)$$

Expanding  $u_h$  in the nodal basis  $\{\phi_j\}_{j=1}^N$ , i.e.,  $u_h(\mathbf{x}, t) = \sum_{j=1}^N c_j(t) \phi_j(\mathbf{x})$ , and choosing  $v = \phi_i$ , we get

$$\sum_{j=1}^N \dot{c}_j \int_{\Omega} \phi_i \phi_j \, d\mathbf{x} + \sum_{j=1}^N c_j \int_{\Omega} \nabla \phi_i \cdot \nabla \phi_j \, d\mathbf{x} = \int_{\Omega} f \phi_i \, d\mathbf{x}, \quad i = 1, 2, \dots, N.$$

This gives a linear system of ODEs for  $\mathbf{c}(t)$

$$M \dot{\mathbf{c}}(t) + A \mathbf{c}(t) = \mathbf{b}(t), \quad t \in [0, T], \quad \mathbf{c}(0) = \mathbf{c}_0,$$

where  $\mathbf{c} = [c_1, c_2, \dots, c_N]^T \in \mathbb{R}^N$  is the unknown vector whose coefficients corresponding to the nodal values of  $u_h$  at a fixed time, and the matrices and vector are given by

$$a_{ij} = \int_{\Omega} \nabla \phi_i \cdot \nabla \phi_j \, d\mathbf{x}, \quad m_{ij} = \int_{\Omega} \phi_i \phi_j \, d\mathbf{x}, \quad b_i = \int_{\Omega} f \phi_i \, d\mathbf{x}, \quad i, j = 1, 2, \dots, N.$$

**Time discretization:** Consider the time discretization of  $[0, T]$  as follows:

$$0 = t_0 < t_1 < \dots < t_M = T$$

with step sizes  $k_m = t_m - t_{m-1}$ ,  $m = 1, 2, \dots, M$ . Let  $\mathbf{c}^{(m)} \approx \mathbf{c}(t_m)$ . If we use the backward Euler method, then the time-stepping scheme reads

$$(M + k_m A) \mathbf{c}^{(m)} = M \mathbf{c}^{(m-1)} + k_m \mathbf{b}_m, \quad m = 1, \dots, M,$$

with the initial vector  $\mathbf{c}^{(0)}$  is computed using the initial condition  $u_0(\mathbf{x}) = u(\mathbf{x}, 0)$ . This can be achieved by direct evaluation at the nodes,  $\mathbf{c}^{(0)} = [u_0(N_1), \dots, u_0(N_N)]^T$ , or by computing the  $L^2$ -projection of  $u_0$  onto  $V_{h,0}$ :

$$\sum_{j=1}^N c_j^{(0)} \int_{\Omega} \phi_i \phi_j \, d\mathbf{x} = \int_{\Omega} u_0 \phi_i \, d\mathbf{x}, \quad i = 1, 2, \dots, N.$$

**Theorem 3.1 (Stability):** The following stability estimates hold for both the continuous and discrete problems:

$$\|u(\cdot, t)\|_{L^2(\Omega)} \leq \|u(\cdot, 0)\|_{L^2(\Omega)} + \int_0^t \|f(\cdot, s)\|_{L^2(\Omega)} ds,$$

$$\|u_h^{(m)}\|_{L^2(\Omega)} \leq \|u_h^{(m-1)}\|_{L^2(\Omega)} + k_m \|f_m\|_{L^2(\Omega)} \leq \|u_h^{(0)}\|_{L^2(\Omega)} + \sum_{i=1}^m k_i \|f_i\|_{L^2(\Omega)}.$$

## 4. The FEM for the wave problem

Many physical systems exhibit behavior that propagates in the form of waves. A familiar example is light, an electromagnetic wave that can spread out (disperse) and

form diffraction patterns-distinct signatures of wave motion. To illustrate the method, we focus on the wave problem defined as follows: determine  $u = u(\mathbf{x}, t)$  satisfying [1]

$$\ddot{u} - \nabla \cdot (c \nabla u) = f, \quad \text{in } \Omega \subset \mathbb{R}^2, \quad t \in [0, T], \quad (23)$$

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \dot{u}(\mathbf{x}, 0) = v_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \quad (24)$$

$$\mathbf{n} \cdot \nabla u(\cdot, t) = 0, \quad \text{on } \partial\Omega, \quad t \in [0, T]. \quad (25)$$

Here,  $f$  is a given source function,  $c = c(\mathbf{x}, t) > 0$  is a material parameter, and the functions  $u_0(\mathbf{x})$  and  $v_0(\mathbf{x})$  are given initial conditions (at  $t = 0$ ). The domain  $\Omega$  is assumed to be bounded region,  $\partial\Omega$  is the boundary of  $\Omega$ , and  $\mathbf{n}$  is its outward unit normal. We want to find a weak solution  $u \in V$ , where the space  $V = H^1(\Omega) = \{v : \|v\| < \infty, \|\nabla v\| < \infty\}$ . Multiplying (23) by a smooth test function  $v$  in  $V$ , integrating over the whole  $\Omega$ , and applying Green's Theorem, we obtain, for  $t \in [0, T]$ ,

$$\begin{aligned} \int_{\Omega} f v \, d\mathbf{x} &= \int_{\Omega} \ddot{u} v \, d\mathbf{x} + \int_{\Omega} c \nabla u \cdot \nabla v \, d\mathbf{x} - \int_{\partial\Omega} c \nabla u \cdot \mathbf{n} v \, ds \\ &= \int_{\Omega} \ddot{u} v \, d\mathbf{x} + \int_{\Omega} c \nabla u \cdot \nabla v \, d\mathbf{x}. \end{aligned}$$

Hence, the variational formulation reads: determine  $u(\cdot, t) \in V$  satisfying [1]

$$\int_{\Omega} \ddot{u} v \, d\mathbf{x} + \int_{\Omega} c \nabla u \cdot \nabla v \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x}, \quad \forall v \in V, \quad t > 0. \quad (26)$$

As before, let  $V_h = \{v \in V : v|_T \in P^1(T), \text{ for } T \in \mathcal{T}_h\} \subset V$  is used to denote the vector space of piecewise continuous linear polynomials on a triangular mesh of the domain  $\Omega$ . Then, the so-called semi-discrete scheme consists of determining  $u_h(\cdot, t) \in V_h$  such that

$$\int_{\Omega} \ddot{u}_h v \, d\mathbf{x} + \int_{\Omega} c \nabla u_h \cdot \nabla v \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x}, \quad \forall v \in V_h. \quad (27)$$

Expanding  $u_h$  in the nodal basis  $\{\phi_j\}_{j=1}^N$ ,  $u_h(\mathbf{x}, t) = \sum_{j=1}^N c_j(t) \phi_j(\mathbf{x})$ , and choosing  $v = \phi_i$  gives

$$\sum_{j=1}^N \ddot{c}_j \int_{\Omega} \phi_i \phi_j \, d\mathbf{x} + \sum_{j=1}^N c_j \int_{\Omega} c \nabla \phi_i \cdot \nabla \phi_j \, d\mathbf{x} = \int_{\Omega} f \phi_i \, d\mathbf{x}, \quad i = 1, 2, \dots, N.$$

This results in the following semi-discrete system of second-order ordinary differential equations (1):

$$M \ddot{\mathbf{c}}(t) + A(t) \mathbf{c}(t) = \mathbf{b}(t), \quad t \in [0, T], \quad (28)$$

where, as before,  $\mathbf{c} = [c_1, c_2, \dots, c_N]^T$  is the vector of nodal unknowns, and the components of the mass matrix  $M \in \mathbb{R}^{N \times N}$  and the stiffness matrix  $A \in \mathbb{R}^{N \times N}$  and the load vector  $\mathbf{b} \in \mathbb{R}^N$  are given by

$$m_{ij} = \int_{\Omega} \phi_i \phi_j d\mathbf{x}, \quad a_{ij} = \int_{\Omega} c \nabla \phi_i \cdot \nabla \phi_j d\mathbf{x}, \quad b_i = \int_{\Omega} f \phi_i d\mathbf{x}, \quad i, j = 1, 2, \dots, N.$$

To solve the system of second-order ODEs, we first reduce it to a first-order linear system of ODEs. Define  $\mathbf{d} = \dot{\mathbf{c}}$ , then

$$M\dot{\mathbf{c}}(t) - M\mathbf{d}(t) = \mathbf{0}, \quad M\dot{\mathbf{d}}(t) + A(t)\mathbf{c}(t) = \mathbf{b}(t), \quad t \in [0, T].$$

Let  $\mathbf{w} = [\mathbf{c}, \mathbf{d}]^T$ . Thus, the linear system of equations can be expressed compactly as

$$\hat{M}\dot{\mathbf{w}}(t) + \hat{A}(t)\mathbf{w}(t) = \hat{\mathbf{b}}(t), \quad t \in [0, T],$$

where

$$\hat{M} = \begin{bmatrix} M & 0 \\ 0 & M \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} 0 & -M \\ A & 0 \end{bmatrix}, \quad \hat{\mathbf{b}} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b} \end{bmatrix}.$$

**The backward Euler method:** Let  $0 = t_0 < t_1 < \dots < t_M = T$  be a discretization of  $[0, T]$  with time steps  $k_m = t_m - t_{m-1}$ . Let  $\mathbf{w}^{(m)} \approx \mathbf{w}(t_m)$ . Using the backward Euler method, we compute

$$(\hat{M} + k_m \hat{A}_m) \mathbf{w}^{(m)} = \hat{M} \mathbf{w}^{(m-1)} + k_m \hat{\mathbf{b}}_m, \quad m = 1, \dots, M,$$

where the initial approximation  $\mathbf{w}^{(0)}$  is determined using the initial conditions  $u_0 = u(\mathbf{x}, 0)$  and  $v_0 = \dot{u}(\mathbf{x}, 0)$ . This can be done by either direct evaluation at the nodes

$$\mathbf{w}^{(0)} = [u_0(N_1), \dots, u_0(N_N), v_0(N_1), \dots, v_0(N_N)]^T,$$

or by computing the  $L^2$ -projection of  $u_0$  and  $v_0$  onto  $V_h$ .

**Crank-Nicolson scheme:** An alternative second-order accurate time-stepping method is.

$$\left( \hat{M} + \frac{k_m}{2} \hat{A}_m \right) \mathbf{w}^{(m)} = \left( \hat{M} - \frac{k_m}{2} \hat{A}_{m-1} \right) \mathbf{w}^{(m-1)} + \frac{k_m}{2} (\hat{\mathbf{b}}_{m-1} + \hat{\mathbf{b}}_m) \equiv \mathbf{g}_m.$$

[Conservation of energy] Suppose that  $f = 0$ . Then the discrete energy is conserved *i.e.*,

$$\| \dot{u}_h(\cdot, t) \|_{L^2(\Omega)}^2 + \| \sqrt{c} \nabla u_h(\cdot, t) \|_{L^2(\Omega)}^2 = \| \dot{u}_h(\cdot, 0) \|_{L^2(\Omega)}^2 + \| \sqrt{c} \nabla u_h(\cdot, 0) \|_{L^2(\Omega)}^2.$$

## 5. The FEM for the Navier-Stokes equations

The motion of a viscous, incompressible fluid is governed by the Navier-Stokes equations, which can be written as follows [31–34]:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p = \mathbf{f}, \quad \text{in } \Omega \times (0, T], \quad (29)$$

$$\nabla \cdot \mathbf{u} = 0, \quad \text{in } \Omega \times (0, T], \quad (30)$$

where  $\mathbf{u}(\mathbf{x}, t)$  is typically the velocity vector field,  $p(\mathbf{x}, t)$  is the pressure,  $\nu > 0$  is the kinematic viscosity,  $\mathbf{f}(\mathbf{x}, t)$  represents external forces, and  $\Omega \subset \mathbb{R}^d$  (with  $d = 2, 3$ )

is the computational domain. Appropriate boundary and initial conditions must also be prescribed. For instance we may impose the following boundary and initial conditions:

**Dirichlet (Velocity) Boundary Condition:** On the portion of the boundary  $\Gamma_D \subset \partial\Omega$ , we prescribe the velocity:  $\mathbf{u}(\mathbf{x}, t) = \mathbf{g}(\mathbf{x}, t)$  for  $\mathbf{x} \in \Gamma_D$  and  $t > 0$ , where the function  $\mathbf{g}$  is prescribed (e.g., no-slip condition  $\mathbf{u} = 0$  on walls).

**Neumann (Stress) Boundary Condition:** On the remaining boundary  $\Gamma_N = \partial\Omega \setminus \Gamma_D$ , we prescribe the traction (stress):  $\mathbf{s}(\mathbf{u}, p) \cdot \mathbf{n} = \mathbf{h}(\mathbf{x}, t)$ ,  $\mathbf{x} \in \Gamma_N$ ,  $t > 0$ , where  $\mathbf{n}$  is the outward unit normal and  $\mathbf{s}(\mathbf{u}, p) = -p\mathbf{I} + \nu(\nabla\mathbf{u} + (\nabla\mathbf{u})^T)$  is the stress tensor.

**Initial Condition:** The velocity field must be initialized at  $t = 0$ :  $\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x})$ ,  $\mathbf{x} \in \Omega$ , where  $\mathbf{u}_0$  satisfies the incompressibility constraint  $\nabla \cdot \mathbf{u}_0 = 0$ .

We first apply the finite element method to discretize the Navier-Stokes equations in space, yielding a system of ODEs in time. This system can then be discretized in time using an ODE solver such as the implicit Euler scheme, a Runge-Kutta method, the Crank-Nicolson method, or a projection method.

By multiplying the momentum Eq. (25) by a test function  $\mathbf{v}$  from the velocity space  $V$ , integrating over the domain  $\Omega$ , and employing integration by parts, we obtain the weak (variational) form [35]:

$$\int_{\Omega} \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v} \, d\mathbf{x} + \int_{\Omega} (\mathbf{u} \cdot \nabla) \mathbf{u} \cdot \mathbf{v} \, d\mathbf{x} + \nu \int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{v} \, d\mathbf{x} - \int_{\Omega} p \nabla \cdot \mathbf{v} \, d\mathbf{x} = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, d\mathbf{x}. \quad (31)$$

For incompressibility, the continuity Eq. (30) is enforced weakly by

$$\int_{\Omega} q \nabla \cdot \mathbf{u} \, d\mathbf{x} = 0, \quad \forall q \in Q, \quad (32)$$

where  $Q$  is the pressure function space.

Let  $V_h \subset V$  and  $Q_h \subset Q$  denote finite-dimensional subspaces for the velocity  $\mathbf{u}$  and pressure  $p$ , respectively. A commonly used stable choice for the pair  $(V_h, Q_h)$  is the Taylor-Hood  $P_2 - P_1$  elements, featuring quadratic approximation for the velocity and linear approximation for the pressure. The resulting semi-discrete finite element formulation is: find  $\mathbf{u}_h(t) \in V_h$  and  $p_h(t) \in Q_h$  such that [35]

$$\int_{\Omega} \frac{\partial \mathbf{u}_h}{\partial t} \cdot \mathbf{v} \, d\mathbf{x} + \int_{\Omega} (\mathbf{u}_h \cdot \nabla) \mathbf{u}_h \cdot \mathbf{v} \, d\mathbf{x} + \nu \int_{\Omega} \nabla \mathbf{u}_h : \nabla \mathbf{v} \, d\mathbf{x} - \int_{\Omega} p_h \nabla \cdot \mathbf{v} \, d\mathbf{x} = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, d\mathbf{x}, \quad (33)$$

$$\int_{\Omega} q \nabla \cdot \mathbf{u}_h \, d\mathbf{x} = 0, \quad (34)$$

for all  $\mathbf{v} \in V_h$  and  $q \in Q_h$ . The FEM approximates the velocity and pressure fields as

$$\mathbf{u} \approx \mathbf{u}_h = \sum_{i=1}^{N_u} \mathbf{U}_i \phi_i(\mathbf{x}), \quad p \approx p_h = \sum_{j=1}^{N_p} P_j \psi_j(\mathbf{x}),$$

where  $\{\phi_i\}$  and  $\{\psi_j\}$  are basis functions for velocity and pressure, and  $\mathbf{U}_i, P_j$  are the nodal coefficients to be determined.

**Time Discretization:** Using a time-stepping scheme such as the backward Euler method, we obtain a fully discrete system at each time step  $t^{n+1}$  [35]:

$$\begin{aligned} \int_{\Omega} \frac{\mathbf{u}_h^{n+1} - \mathbf{u}_h^n}{\Delta t} \cdot \mathbf{v} \, d\mathbf{x} + \int_{\Omega} (\mathbf{u}_h^{n+1} \cdot \nabla) \mathbf{u}_h^{n+1} \cdot \mathbf{v} \, d\mathbf{x} + \nu \int_{\Omega} \nabla \mathbf{u}_h^{n+1} : \nabla \mathbf{v} \, d\mathbf{x} - \int_{\Omega} p_h^{n+1} \nabla \cdot \mathbf{v} \, d\mathbf{x} \\ = \int_{\Omega} \mathbf{f}^{n+1} \cdot \mathbf{v} \, d\mathbf{x}, \end{aligned} \quad (35)$$

$$\int_{\Omega} q \nabla \cdot \mathbf{u}_h^{n+1} \, d\mathbf{x} = 0. \quad (36)$$

## 6. Remarks

The finite element spaces  $(V_h, Q_h)$  must satisfy the *inf-sup* (or Ladyzhenskaya-Babuška-Brezzi) condition to guarantee stability of the mixed formulation (see Brezzi and Fortin [36]).

- Nonlinear convective terms often require iterative solvers, such as Newton's method or Picard iteration, at each time step.
- The semi-discrete finite element formulation results in a system of ODEs in time, which can subsequently be integrated using standard time-stepping schemes.

## 7. Conclusion

In this chapter, we presented the core concepts of the classical FEM and illustrated its application to approximating solutions of various PDEs. We started with second-order elliptic problems in two dimensions and then extended the discussion to time-dependent equations, including the diffusion (heat) equation, the convection-diffusion equation, the wave equation, and the incompressible Navier-Stokes equations. The goal was to establish a solid foundation for understanding how FEM transforms differential equations into discrete systems that can be solved numerically. In the chapters that follow, we will build on these principles and apply the method to more advanced topics, particularly in elasticity and structural mechanics, where elliptic PDEs and complex geometries are central. For readers interested in exploring the theoretical background or more specialized applications, comprehensive treatments can be found in [1, 3–10] and the references therein.

## **Author details**

Mahboub Baccouch

Department of Mathematics and Statistical Sciences, University of Nebraska at Omaha, Omaha, USA

\*Address all correspondence to: [mbaccouch@unomaha.edu](mailto:mbaccouch@unomaha.edu)

## **IntechOpen**

---

© 2026 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Baccouch M. Finite Element Methods and Their Applications. London: IntechOpen; 2021
- [2] Gockenbach MS. Understanding and Implementing the Finite Element Method. United States: SIAM; 2006
- [3] Ainsworth M, Oden JT. A Posteriori error estimation in Finite Element Analysis. New York: John Wiley; 2000
- [4] Brenner SC, Scott LR. The mathematical theory of Finite Element Methods. 3 ed. New York: Springer-Verlag; 2008
- [5] Ciarlet PG. The Finite Element Method for Elliptic Problems. Amsterdam-New York-Oxford: North-Holland Pub. Co.; 1978
- [6] Johnson C. Numerical solution of partial differential equations by the Finite Element Method. New York: Cambridge University Press; 1987
- [7] Larson M. The Finite Element Method: Theory, implementation, and applications. Berlin New York: Springer; 2013
- [8] Oden JT, Carey GF. Finite Elements, Mathematical Aspects. Englewood Cliffs: Prentice Hall; 1983
- [9] Schwab C. P- and Hp- Finite Element Methods. New York: Oxford University Press; 1998
- [10] Szabo B, Babuška I. Finite Element Analysis. New York: Wiley; 1991
- [11] Kaltenbacher M. Numerical simulation of mechatronic sensors and actuators: Finite elements for computational multiphysics. Heidelberg: Springer; 2015
- [12] Hrennikoff A. Solution of problems of elasticity by the framework method. Journal of Applied Mechanics. 1941;8 (4):169–175
- [13] Courant R. Variational methods for the solution of problems of equilibrium and vibrations. Bulletin Of The American Mathematical Society. 1943;49:1–23
- [14] Clough RW. The finite element method in plane stress analysis. In Proceedings of the Second ASCE Conference on Electronic Computation, 345–378, Pittsburgh, 1960
- [15] Zienkiewicz OC, Taylor RL. The Finite Element Method: Volume 1: The Basis. 5th ed. United States: Butterworth–Heinemann; 2000
- [16] Oden JT. Finite Elements of Nonlinear Continua. United States: McGraw–Hill; 1972
- [17] Strang G, Fix GJ. An Analysis of the Finite Element Method. United States. 1973
- [18] Babuška I, Aziz AK. Survey lectures on the mathematical foundations of the finite element method. In The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations. United States: Academic Press; 1972
- [19] Brezzi F, Fortin M. Mixed and Hybrid Finite Element Methods. New York, USA: Springer; 1991

- [20] Babuška I, Rheinboldt WC. Error estimates for adaptive finite element computations. *SIAM Journal on Numerical Analysis*. 1978;15:736–754
- [21] Verfürth R. A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques. UK: Wiley-Teubner; 1996
- [22] Arnold DN, Brezzi F, Cockburn B, Marini LD. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM Journal on Numerical Analysis*. 2002;39(5): 1749–1779
- [23] Cockburn B, Shu C-W. Runge–Kutta discontinuous Galerkin methods for convection-dominated problems. *Journal of Scientific Computing*. 2001;16:173–261
- [24] Hughes TJR, Cottrell JA, Bazilevs Y. Isogeometric analysis: CAD, finite elements, nurbs, exact geometry, and mesh refinement. *Computer Methods in Applied Mechanics and Engineering*. 2005;194:4135–4195
- [25] Ghanem RG, Spanos PD. *Stochastic Finite Elements: A Spectral Approach*. New York, NY: Springer; 1991
- [26] Canuto C, Hussaini MY, Quarteroni A, Zang TA. *Spectral Methods: Fundamentals in Single Domains*. New York, NY: Springer; 2007
- [27] Quarteroni A, Valli A. *Domain decomposition methods for Partial differential equations*. Oxford, United Kingdom: Oxford University Press; 1999
- [28] Ern A, Guermond J-L. *Theory and Practice of Finite Elements*. New York, USA: Springer; 2004
- [29] Hughes TJR. *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*. New Jersey, USA: Prentice Hall; 1987
- [30] Ern A, Guermond J-L. *Theory and Practice of Finite Elements*. Vol. 159. New York, NY: Springer; 2004
- [31] Batchelor GK. *An Introduction to Fluid Dynamics* England. Cambridge University Press; 1967
- [32] Ladyzhenskaya OA. *The Mathematical Theory of Viscous Incompressible Flow*. New York, USA: Gordon and Breach; 1969
- [33] Temam R. *Navier-Stokes Equations: Theory and Numerical Analysis*. Amsterdam: North-Holland; 1977
- [34] Lions P-L. *Mathematical Topics in Fluid Mechanics, Volume 1: Incompressible Models*. Cambridge, UK: Oxford University Press; 1996
- [35] Temam R. *Navier–Stokes Equations: Theory and Numerical Analysis*. Providence, RI: AMS Chelsea Publishing; 2001
- [36] Brezzi F, Fortin M. *Mixed and Hybrid Finite Element Methods*. In of *Springer Series in Computational Mathematics*. Vol. 15. New York, NY: Springer-Verlag; 1991. p. 359



---

## Section 2

# Classical and Advanced FEM Formulations

---



## Chapter 2

# The Classical Galerkin Finite Element Method Applied to Neutron Diffusion Theory

*Lenilson Moreira Araujo and Anderson Alvarenga de Moura Meneses*

### Abstract

This chapter aims to present a numerical strategy to solve the neutron diffusion equation (NDE) using the classical finite element method, also known as the Galerkin finite element method (GFEM). To this end, a particular case of the NDE in steady-state and two-dimensional geometry is presented. The mathematical development of the variational formulation of the NDE, when discretized by the GFEM, is then shown. A brief description of the GFEM implementation and a GFEM flowchart for the NDE considering two energy groups are provided. Two benchmark problems are solved using bilinear shape functions for triangular elements and quadrilateral elements. The results corroborate the good convergence rate of the GFEM when compared to the finite difference method (FDM).

**Keywords:** classical Galerkin, variational formulation, triangular elements, quadrilateral elements, accuracy analysis, reactor physics, neutron diffusion equation

### 1. Introduction

Nuclear power reactors operate through the fission of specific isotopes of uranium and plutonium, triggered by neutrons at certain energy levels. There are several concepts of nuclear power reactors, one of which is the pressurized water reactor (PWR) [1]. To control the fission rate, manage the neutron population, and prevent the spread of radiation, the core of nuclear reactors contains several essential structures, which are described as follows. Fuel assemblies (FAs): These structures house the nuclear fuel. Control rods: Made of materials with high neutron absorption capacity and these rods regulate the fission rate. Coolant: These are primarily used to remove heat, but can also act as a neutron moderator, reducing neutron energy levels. Reflectors: These components redirect neutrons escaping from the reactor core's periphery back into the core. Shielding: Designed to contain all radiation generated within the reactor core. To simplify the calculation of the neutron population inside the reactor core, the design often incorporates 1/2, 1/4, and 1/8 symmetry. **Figure 1** illustrates the

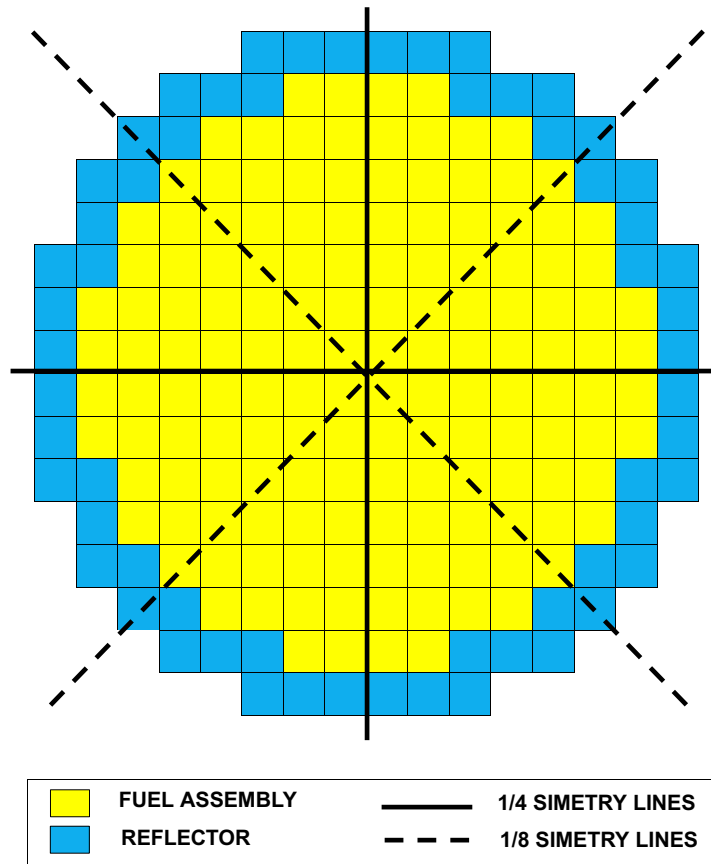


**Figure 1.**  
Top view of a PWR core.

two-dimensional (2D) modeling of the reactor, as if viewed from above, with each yellow square representing an FA. **Figure 2** provides a detailed view of an FA.

One of the fundamental roles of reactor physics is to design nuclear reactors capable of maintaining fission reactions in a self-sustainable and safe manner. For this purpose, the main parameters are the neutron flux  $\Phi_g$  and the effective neutron multiplication factor  $k_{eff}$ . The first corresponds to the neutron population within the reactor core; the second, as the ratio between the number of neutrons produced and the number of neutrons absorbed, indicates the criticality condition of the reactor [2]. If  $k_{eff} < 1$ , the reactor is said to be subcritical. If  $k_{eff} = 1$ , the reactor is said to be critical and if  $k_{eff} > 1$ , the reactor is said to be supercritical. In general, the supercritical PWR reactor is designed, and the criticality condition is reached using control rods or soluble Boro.

In the process of designing nuclear reactors, it is essential to obtain microscopic cross sections that represent with high fidelity the physical reactions that occur inside the reactor and that keep it operating properly. These parameters are obtained through the neutron transport equation (NTE), which is a first-order hyperbolic partial differential equation (PDE). Once the microscopic cross sections are generated, the macroscopic cross sections are obtained, which can be homogenized,

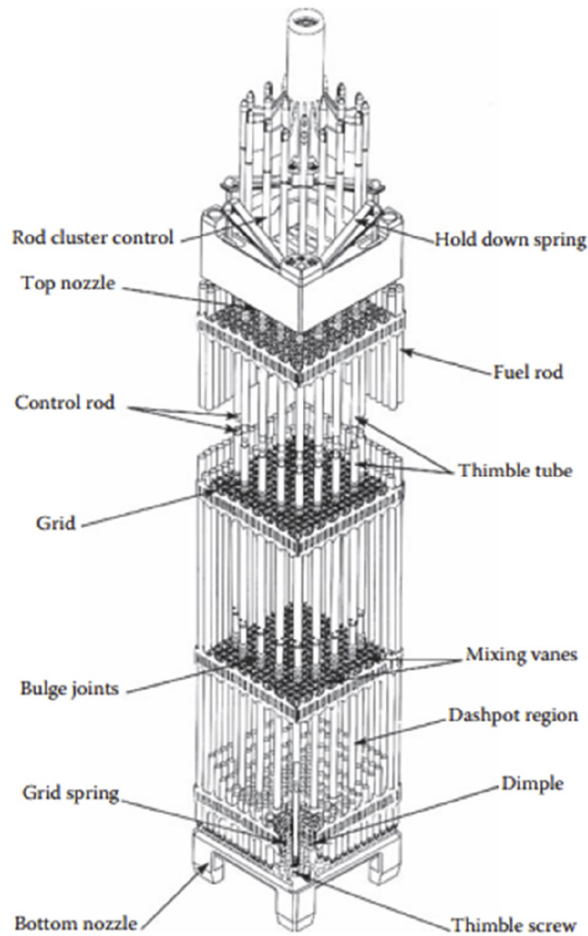


**Figure 2.**  
 Typical fuel assembly for a PWR [1].

considering the different types of materials that make up the reactor core, in order to simplify its physical composition with the aim of reducing the computational effort required to determine the parameters  $\Phi_g$  and  $k_{eff}$ . To further reduce computational effort, in many situations, such as nuclear reactor design analysis, an approximation of the NTE known as neutron diffusion equation (NDE) is used, which, being mathematically and physically simpler, requires much less computational effort to solve [3].

The neutron flux in theoretical one-dimensional (1D) and 2D reactors essentially preserves the same behavior observed in three-dimensional (3D) reactors; that is, the flux is higher in the center of the reactor core and lower at the periphery. This fact makes it acceptable to perform a preliminary analysis of the reactor core in 1D or 2D before it is designed in 3D. In this chapter, prioritizing a more didactic approach, we will focus on the criticality and power analysis of two-dimensional reactors with homogenized nuclear parameters. Several numerical methods are used to solve the NDE, among which we can mention the finite difference method (FDM) [4], the nodal expansion method (NEM) [5, 6], analytical nodal methods (ANMs) [7–9],

and several methods from the finite element method (FEM) family [10]. In this chapter, we adopt the Galerkin FEM (GFEM) [10, 11] to solve a particular case of the NDE. Below, we will briefly review the history of the FEM.

Having emerged in the middle of the last century in structural engineering, FEM has been increasingly used in several areas of engineering. This is mainly due to its ability to deal with non-regular geometries and unstructured meshes, in addition to the fact that it presents excellent convergence rates. Such characteristics make FEM very attractive when compared to its most common alternative, the FDM. As an example of the several areas in which the application of FEM has grown in recent decades, we can mention reactor physics. Much research has been carried out focusing on developing computational codes based on FEM, exploring its ability to deal with various types of geometry [10, 12]. Other research has sought to optimize FEM calculation time while preserving its accuracy [12, 13].

This chapter focuses on reactor physics and aims to present a numerical strategy for solving the NDE using GFEM. To this end, we will first introduce a particular case of the NDE in steady-state and two-dimensional geometry. Subsequently, we will detail the mathematical development of the variational formulation of the NDE when discretized by the GFEM. Next, a brief description of the GFEM implementation will be provided. The following step involves presenting the problems to be solved, followed by the presentation and discussion of the results. Finally, the conclusion will be drawn along with the final considerations.

## 2. Two-dimensional multigroup NDE in steady state

The NDE is an elliptic PDE that describes the variation in neutron density in a control volume  $\Omega$  and represents the balance between neutron loss and production within that volume.

Let the spatial domain be denoted by  $\Omega = \{(x, y) \in \mathbb{R}^2\}$ , the strong formulation of the NDE for energy multigroup, in steady-state and in 2D Cartesian geometry, is given by

$$-\nabla \cdot D_g(x, y) \nabla \Phi_g(x, y) + \Sigma_{r,g}(x, y) \Phi_g(x, y) = Q_{S,g}(x, y) + Q_{F,g}(x, y) \quad (1)$$

$$g = 1, \dots, G,$$

$$\forall (x, y) \in \Omega \subset \mathbb{R}^2,$$

where  $Q_{S,g}$  is the scattering source term defined as

$$Q_{S,g}(x, y) = \sum_{g' \neq g}^G \Sigma_{s,g',g}(x, y) \Phi_{g'}(x, y), \quad (2)$$

and  $Q_{F,g}$  is the fission source term, which for multiplicative media is given as

$$Q_{F,g}(x, y) = \frac{\chi_g}{k_{eff}} \sum_{g'=1}^G \nu \Sigma_{f,g'}(x, y) \Phi_{g'}(x, y). \quad (3)$$

For Equations (1), (2) and (3), we have the nuclear parameters defined in the usual way [2], as follows:

$D_g(x, y)$  is the diffusion coefficient in group  $g$ ;

$\Phi_g(x, y)$  is the neutrons flux in group  $g$ ;

$\Sigma_{r,g}(x, y)$  is the removal macroscopic cross section in group  $g$ ;

$\Sigma_{s,g',g}(x, y)$  is the scattering macroscopic cross section from group  $g'$  to group  $g$ ;

$\Sigma_{f,g'}(x, y)$  is the fission macroscopic cross section in group  $g'$ ;

$\nu$  is the average number of neutrons released in a fission reaction induced by a neutron in group  $g'$ ;

$\chi_g$  is the fission spectrum and means the probability that a fission neutron will be born with an energy in group  $g$ ;

$k_{eff}$  is the neutron effective multiplication factor.

In a nuclear power reactor, neutrons originating from fission are born with energies of up to 10 MeV and can be moderated to 0.001 eV. The macroscopic cross sections present in the NDE depend directly on the energy of the incident neutron. For this reason, the energy variable is discretized into intervals, or groups, with the aim of simplifying and standardizing the energy dependence of the reactions that occur within a nuclear reactor.

The discretization of the energy variable is performed according to the scheme shown in **Figure 3(a)**, considering that the neutron loses energy throughout its lifetime. For two energy groups, the structure is shown in **Figure 3(b)**.

In this chapter, we use two types of boundary conditions. The first is the vacuum boundary condition or zero reentrant current condition, which specifies that there are no neutrons entering the control volume and is given as follows:

$$J_g^{in}|_{\partial\Omega} = \frac{\Phi_g(x, y)|_{\partial\Omega}}{4} + \frac{D}{2} \frac{\partial \Phi_g(x, y)}{\partial n} |_{\partial\Omega} = 0, \quad (4)$$

where

$$\frac{\partial \Phi_g(x, y)}{\partial n} = \nabla \Phi_g(x, y) \cdot \vec{n}, \quad (5)$$

and  $\vec{n}$  is the unit vector normal to the surface  $\partial\Omega$ . The second is a reflective boundary condition that establishes that the number of neutrons leaving the control

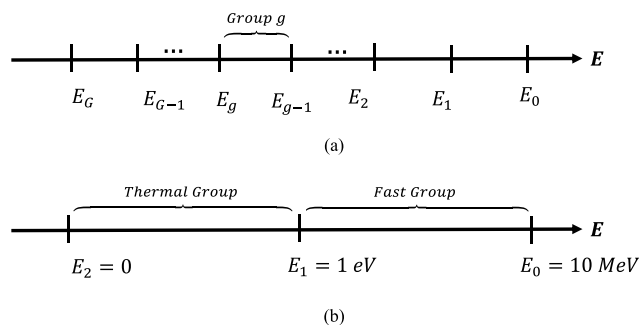


**Figure 3.**  
 (a) Discretization of the energy variable for  $G$  energy groups. (b) Discretization of the energy variable for two energy groups.

volume through the surface  $\partial\Omega$  is equal to the number of neutrons reaching the control volume through that same surface and is given as follows:

$$J_{net}|_{\partial\Omega} = -D_g(x, y) \frac{\partial\Phi_g(x, y)}{\partial n} = 0 \Rightarrow \frac{\partial\Phi_g(x, y)}{\partial n} = 0. \quad (6)$$

### 3. Discretization of the multigroup NDE with GFEM

In order to present the finite element formulation to be used for the spatial discretization of the problem defined by Equation (1), we will introduce the variational formulation of the NDE with energy multigroup.

Consider  $\Omega$  as an open set of  $\mathbb{R}^2$ , let  $S(\Omega)$  be the space where the solution  $\Phi_g$  is found and  $V(\Omega)$  be the space of admissible variations where the projection functions  $w$  are found. The Galerkin variational formulation of the NDE can be defined as follows:

Seek  $\Phi_g \in S(\Omega)$ , such that

$$A(\Phi_g, w) = F(w), \quad \forall w \in V(\Omega), \quad (7)$$

where

$$\begin{aligned} A(\Phi_g, w) = & \int_{\Omega} D_g(x, y) \nabla \Phi_g(x, y) \nabla w(x, y) d\Omega - \int_{\partial\Omega} D_g(x, y) \frac{\partial\Phi_g(x, y)}{\partial n} w(x, y) ds + \\ & + \int_{\Omega} \Sigma_{r,g}(x, y) \Phi_g(x, y) w(x, y) d\Omega \end{aligned} \quad (8)$$

and

$$F(w) = \int_{\Omega} [Q_{S,g}(x, y) + Q_{F,g}(x, y)] w(x, y) d\Omega. \quad (9)$$

The bilinear form and functional are defined by Equations (8) and (9) and, respectively, are obtained by multiplying Equation (1) by the projection function  $w(x, y) \in V(\Omega)$ , integrating over the entire spatial domain  $\Omega$  and using Green's Theorem on the differential term.

When applying the FEM to Equations (8) and (9), we proceed as follows:

- The domain  $\Omega$  is partitioned as a set  $\{\Omega_1, \dots, \Omega_{Ne}\}$  consisting of  $Ne$  elements  $\Omega_e$  which are non-degenerate and disjoint, that is,  $\Omega_i \cap \Omega_j = \emptyset$  if  $i \neq j$ . Then we replace the integrals  $\int_{\Omega} d\Omega$  in the interior of the domain and on its boundary  $\int_{\partial\Omega} ds$  by the sum of integrals  $\sum_{e=1}^{Ne} \int_{\Omega_e} d\Omega_e$  and  $\sum_{e=1}^{Ne} \int_{\partial\Omega_e} ds_e$ , respectively;
- The neutron flux  $\Phi_g$  is established as a linear combination of the form  $\Phi_g^e(x, y) = \sum_{i=1}^{Npel} N^{e,i}(x, y) \phi_g^{e,i}$ , where  $\Phi_g^e$  is the restriction of the function  $\Phi_g$  to the element  $\Omega_e$ . For each element  $\Omega_e$ ,  $Npel$  is the number of vertices of the element  $\Omega_e$  and each  $\phi_g^i$  corresponds to numeric value of the flux  $\Phi_g^e$  at the

vertex  $i$  of the element  $\Omega_e$ . For each element  $\Omega_e$ , the functions  $N^{e,i}$  are the components of the base function  $N^e(x, y) = \sum_{i=1}^{N_{pel}} N^{e,i}(x, y)$ .

For the classical Galerkin formulations, the spaces  $S(\Omega)$  and  $V(\Omega)$  are coincident, so that  $(\Omega) = V(\Omega) = H^1(\Omega)$ , where  $H^1(\Omega)$  is the usual Sobolev space as defined in the study by Adams [14], as:

$$H^1(\Omega) = \left\{ u : \Omega \rightarrow \mathbb{R}; u \in L^2(\Omega), \frac{\partial u}{\partial x_i} \in L^2(\Omega), i = 1, 2 \right\} \quad (10)$$

where

$$L^2(\Omega) = \left\{ u : \Omega \rightarrow \mathbb{R}; \int_{\Omega} u^2 d\Omega < \infty \right\} \quad (11)$$

Thus, for each element  $\Omega_e$  the projection functions  $w(x, y) \in V(\Omega)$  and the base functions  $N^e(x, y) \in S(\Omega)$  coincide.

With these considerations, one can finally present the special finite element discretization of the problem described by Equation (1) as follows:

Seek  $\Phi_g \in S^h(\Omega) \subset H^1(\Omega)$ , such that

$$\tilde{A}(\Phi_g, w) = \tilde{F}(w), \forall w \in V^h(\Omega) \subset H^1(\Omega). \quad (12)$$

$S^h$  and  $V^h$  are finite element spaces and the bilinear forms as well as the functionals of the Galerkin method are defined by

$$\tilde{A}(\Phi_g, w) = \sum_{e=1}^{N_e} \left[ \int_{\Omega_e} D_g \nabla N^{eT} \Phi_g^e \nabla N^e d\Omega_e - \int_{\partial\Omega_e} \frac{\Phi_g^e}{2} N^e ds_e + \int_{\Omega_e} \Sigma_{r,g}^e N^{eT} \Phi_g^e N^e d\Omega_e \right] \quad (13)$$

and

$$\tilde{F}(w) = \sum_{e=1}^{N_e} \int_{\Omega_e} \left[ \frac{\chi_g}{k_{eff}} \sum_{g'=1}^G \nu \Sigma_{f,g'}^e N^{eT} \Phi_{g'}^e + \sum_{g' \neq g} \Sigma_{s,g',g}^e N^{eT} \Phi_{g'}^e \right] N^e d\Omega_e. \quad (14)$$

### 3.1 GFEM implementation

In this chapter, triangular bilinear elements and rectangular bilinear elements are used to implement GFEM. The local matrices are defined by

$$A_{ij}^e X_i^e = \int_{\Omega_e} D_g^e \nabla N^{e,i} \phi_g^{e,i} \nabla N^{e,j} d\Omega_e + \int_{\Omega_e} \Sigma_{r,g}^e N^{e,i} \phi_g^{e,i} N^{e,j} d\Omega_e \quad (15)$$

and

$$F_i^e = \int_{\Omega_e} \left( \frac{\chi_g}{k_{eff}} \sum_{g'=1}^G \nu \Sigma_{f,g'}^e N^{e,i} \phi_{g'}^{e,i} + \sum_{g' \neq g} \Sigma_{s,g',g}^e N^{e,i} \phi_{g'}^{e,i} \right) N^{e,j} d\Omega_e + \int_{\partial\Omega_e} D_g \frac{\partial \phi_g^{e,i}}{\partial n} N^{e,j} ds_e. \quad (16)$$

In Equations (15) and (16), the element-level integrals  $\Omega_e$  are computed through Gaussian quadrature using the change of variables known as isoparametric

mapping. The integrals presented in Equations (15) and (16) are reorganized to form the local matrices  $A_{i,j}^e X_i^e$  and  $F_i^e$ . Then, through the assembly process [15], the local matrices are structured to form the global matrices  $\tilde{A}$  and  $\tilde{F}$ . Such procedures result in a large-scale linear algebraic system of the form  $AX = F$ . It is important to highlight that the global matrix  $A$  originated from GFEM is positive, symmetric, and sparse, which makes it possible to use solvers based on the LDU decomposition. In this chapter, we use a particular case of the LDU decomposition known as Crout's method [16]. The algorithm flowchart for the numerical solution of NDE by means of GFEM is shown in **Figure 4**.

For each iteration  $n$  the fission source is updated through Equation (17) as follows:

$$Q_F^n(x, y) = \frac{\chi_g}{k_{eff}^n} \sum_{g=1}^G \nu \Sigma_{f,g'}(x, y) \phi_g^n(x, y). \quad (17)$$

In turn, the eigenvalue  $k_{eff}$  is updated through Equation (18) as follows:

$$k_{eff}^n = k_{eff}^{n-1} \frac{\int_{\Omega} Q_F^n(x, y) d\Omega}{\int_{\Omega} Q_F^{n-1}(x, y) d\Omega}. \quad (18)$$

### 3.1.1. Linear basis functions for triangular elements

When using triangular elements, the element  $\Omega_e(x, y)$  is parameterized in an isosceles right triangle  $\hat{\Omega}_e(\xi, \eta)$  according to **Figure 5**. For bilinear triangular elements, the basis functions are defined as follows:

$$N^{e,i}(\xi, \eta) = \frac{a_i + b_i \xi + c_i \eta}{2\Delta^e}, \quad i = 1, 2, 3, \quad (19)$$

where

$$\begin{aligned} a_1 &= (\xi_2 \eta_3 - \xi_3 \eta_2), \quad a_2 = (\xi_3 \eta_1 - \xi_1 \eta_3), \quad a_3 = (\xi_1 \eta_2 - \xi_2 \eta_1), \\ b_1 &= (\eta_2 - \eta_3), \quad b_2 = (\eta_3 - \eta_1), \quad b_3 = (\eta_1 - \eta_2), \\ c_1 &= (\xi_3 - \xi_2), \quad c_2 = (\xi_1 - \xi_3), \quad c_3 = (\xi_2 - \xi_1) \end{aligned}$$

and

$$2\Delta^e = \begin{vmatrix} 1 & \xi_1 & \eta_1 \\ 1 & \xi_2 & \eta_2 \\ 1 & \xi_3 & \eta_3 \end{vmatrix}.$$

### 3.1.2. Linear basis functions for quadrilateral elements

When using quadrilateral elements, the element  $\Omega_e(x, y)$  is parameterized in a biunitary square  $\hat{\Omega}_e(\xi, \eta) \in [-1, 1] \times [-1, 1]$ , according to **Figure 6**. For bilinear quadrilateral elements, the basis functions are defined as follows:

$$N^{e,i}(\xi, \eta) = \frac{1}{4} (1 + \xi_i \xi) (1 + \eta_i \eta), \quad i = 1, 2, 3, 4. \quad (20)$$

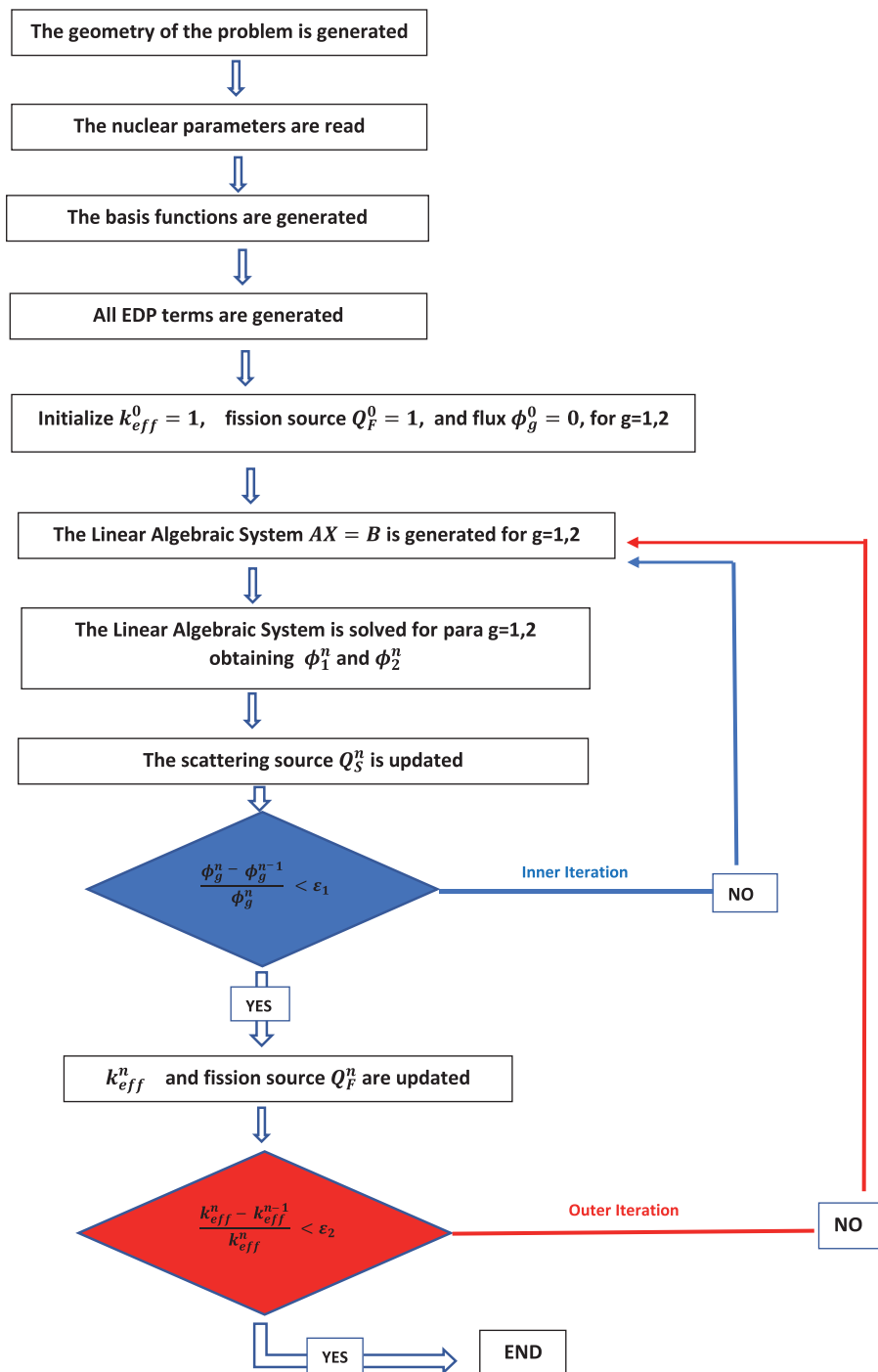


**Figure 4.**  
 GFEM flowchart for the two-energy group neutron diffusion problem.

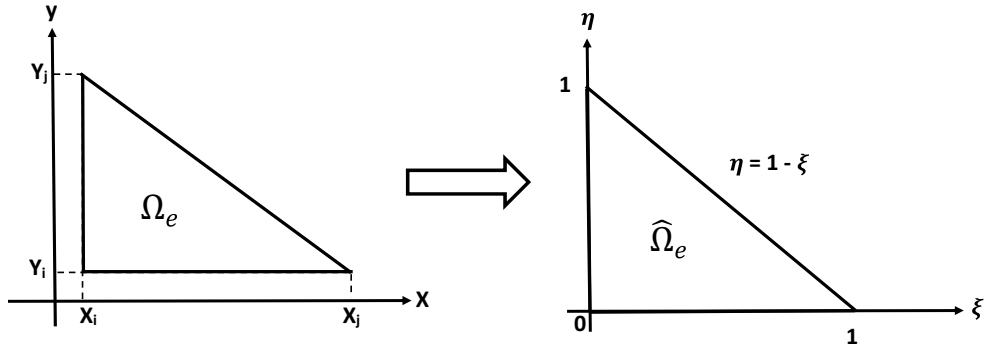


**Figure 5.**  
Simplified isoparametric mapping scheme for triangular elements.

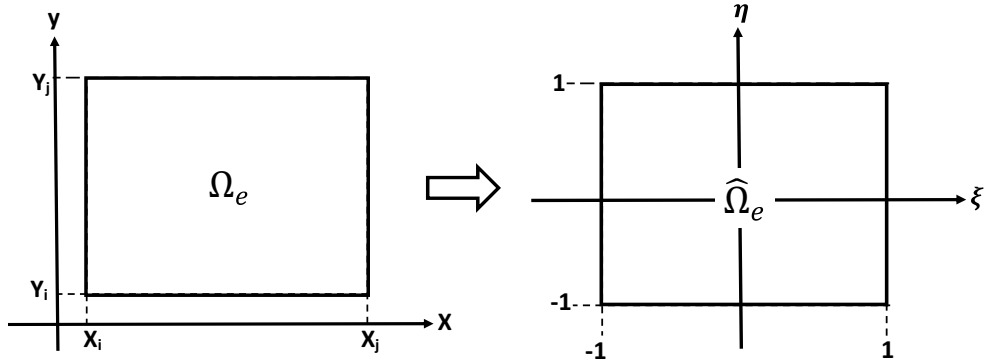


**Figure 6.**  
Simplified isoparametric mapping scheme for quadrilateral elements.

## 4. Description of the problems

### 4.1 Homogeneous reactor 2D

The homogeneous reactor benchmark [17] is a square with dimensions of  $200 \times 200 \text{ cm}^2$ . The problem is defined with two energy groups, and the nuclear data are presented in **Table 1**. This problem considers the condition of zero flux along the entire external boundary of the reactor and the condition of zero net current under the symmetry axes of  $\frac{1}{4}$ , as shown in **Figure 7**.

### 4.2 EPRI-9 2D reactor

This is a benchmark from a small PWR. Originally, it has eight FAs measuring  $21 \times 21 \text{ cm}^2$  surrounded by a 2.8 cm thick baffle and water reflector, without control

| $g$ | $D_g(\text{cm})$ | $\Sigma_{a,g}(\text{cm}^{-1})$ | $\Sigma_{s,g' \rightarrow g}(\text{cm}^{-1})$ | $\Sigma_{f,g}(\text{cm}^{-1})$ | $\chi_g$ |
|-----|------------------|--------------------------------|-----------------------------------------------|--------------------------------|----------|
| 1   | 1.35             | 0.001382                       | 0.0000                                        | 0.0005832                      | 1.0      |
| 2   | 1.08             | 0.005690                       | 0.0023                                        | 0.0103148                      | 0.0      |

**Table 1.**  
*Nuclear parameters for a homogeneous reactor.*

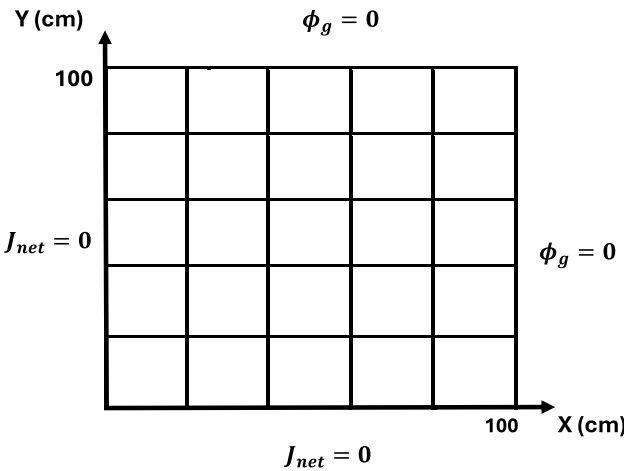


**Figure 7.**  
*A quarter of the homogeneous reactor 2D benchmark.*

| Region | $g$ | $D_g(\text{cm})$ | $\Sigma_{a,g}(\text{cm}^{-1})$ | $\Sigma_{s,g' \rightarrow g}(\text{cm}^{-1})$ | $\Sigma_{f,g}(\text{cm}^{-1})$ | $\chi_g$ |
|--------|-----|------------------|--------------------------------|-----------------------------------------------|--------------------------------|----------|
| 1      | 1   | 1.5089E + 00     | 1.2063E-02                     | 0.0                                           | 5.9942E-03                     | 1.0      |
|        | 2   | 4.2521E-01       | 1.8142E-01                     | 2.1063E-02                                    | 2.3556E-01                     | 0.0      |
| 2      | 1   | 1.5079E + 00     | 9.2914E-03                     | 0.0                                           | 4.6083E-03                     | 1.0      |
|        | 2   | 4.2105E-01       | 1.5067E-01                     | 2.1049E-02                                    | 1.7531E-01                     | 0.0      |
| 3      | 1   | 1.1635E + 00     | 1.5955E-03                     | 0.0                                           | 0.0                            | 0.0      |
|        | 2   | 1.2099E + 00     | 2.1002E-01                     | 1.8462E-02                                    | 0.0                            | 0.0      |

**Table 2.**  
*EPRI-9 2D homogeneous nuclear parameters.*

rods inserted, and two types of fuel FAs (C1 and C2) containing different enrichments. However, in this chapter, we are considering homogenized nuclear parameters where the baffle and water reflectors constitute Region 3, and there are no water holes inside the FAs, as in the study by Pessoa et al. [18]. The nuclear parameters are presented in **Table 2**, and **Figure 8** shows one-quarter of the geometry of this reactor as well as the boundary conditions used.

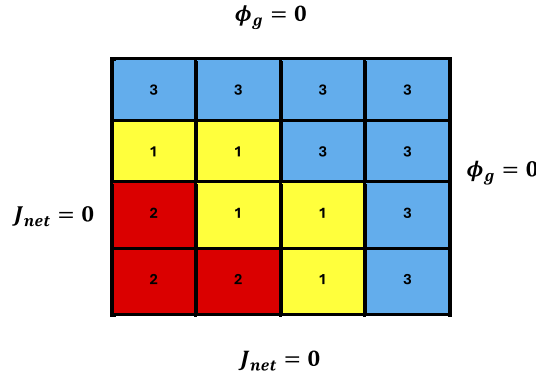


**Figure 8.**  
One-quarter of the EPRI-9 2D benchmark.

## 5. Results and discussions

All simulations were performed using the convergence parameters  $\varepsilon_1 = 10^{-6}$  and  $\varepsilon_2 = 10^{-9}$ . For GFEM, the triangular mesh grid was generated from rectangular mesh grid by inserting one diagonal in each rectangular element. The reference values were obtained through the FDM with  $100 \times 100$  partitions for the homogeneous reactor and  $120 \times 120$  partitions for the EPRI-9 reactor. The powers presented were calculated through Equation (21) as follows:

$$P = \sum_{g=1}^G \nu \Sigma_{f,g}(x, y) \Phi_g(x, y). \quad (21)$$

In turn, the percentage relative errors presented were obtained through Equation (22) as follows:

$$PRE(\%) = \left( \frac{\text{reference value} - \text{obtained value}}{\text{reference value}} \right) \times 100\%. \quad (22)$$

## 6. Homogenous reactor 2D

**Figure 9** shows the neutron flux for  $g = 1$  (fast flux) and  $g = 2$  (thermal flux) for the homogeneous reactor. **Table 3** shows the convergence of the effective neutron multiplication factor. **Table 4** shows the power convergence. **Figure 10** shows the reference power values obtained with the FDM and the percentage relative errors calculated by comparing them with the power obtained with the GFEM using  $50 \times 50$  partitions. The normalized average power in each FA was calculated considering that each FA has dimensions  $20 \times 20 \text{ cm}^2$ .

For the homogeneous reactor problem, in all simulations, the bilinear rectangular elements presented better convergence in both the  $k_{eff}$  eigenvalue analysis and the power analysis. For the bilinear triangular elements, the smallest error presented for

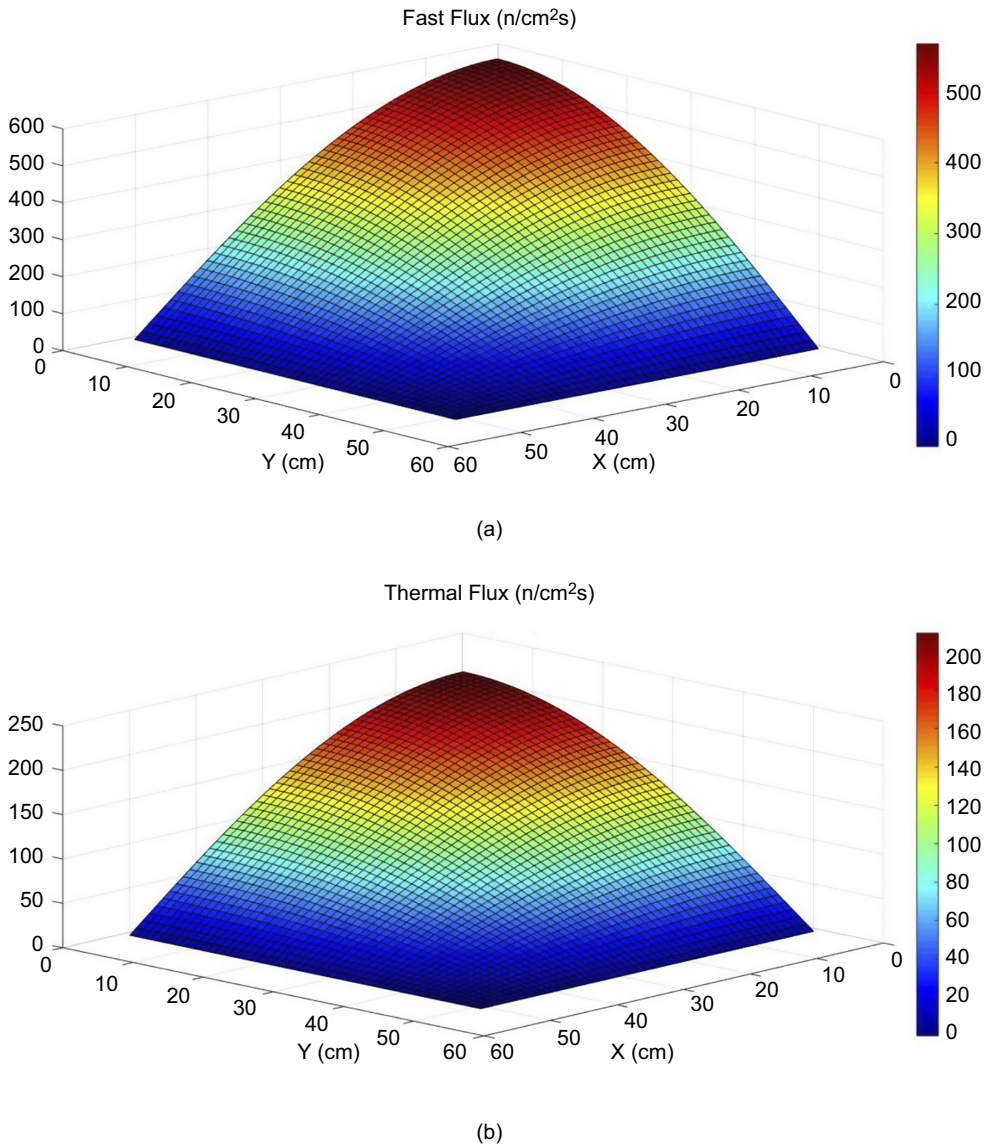


**Figure 9.** Homogeneous reactor 2D (a) fast neutron flux and (b) thermal neutron flux obtained from the GFEM method with  $50 \times 50$  partitions.

the eigenvalue and power are 0.015163 and 0.035616, respectively. In turn, considering the bilinear rectangular elements, the smallest error presented for the eigenvalue and power are 0.003961 and 0.058826, respectively. It is worth mentioning that when it comes to power distribution, even for the lowest discretization level,  $05 \times 05$ , the bilinear rectangular elements present an error of 0.058826.

| Triangular elements            |           |          | Quadrilateral elements |           |          |
|--------------------------------|-----------|----------|------------------------|-----------|----------|
| Number of elements             | $k_{eff}$ | PRE(%)   | Number of elements     | $k_{eff}$ | PRE(%)   |
| 05 × 05 × 02                   | 1.006802  | 0.406187 | 05 × 05                | 1.008993  | 0.189521 |
| 10 × 10 × 02                   | 1.009776  | 0.112009 | 10 × 10                | 1.010414  | 0.048887 |
| 15 × 15 × 02                   | 1.010341  | 0.056092 | 15 × 15                | 1.010677  | 0.022883 |
| 20 × 20 × 02                   | 1.010540  | 0.03642  | 20 × 20                | 1.010769  | 0.013785 |
| 40 × 40 × 02                   | 1.010732  | 0.017434 | 40 × 40                | 1.010858  | 0.005014 |
| 50 × 50 × 02                   | 1.010755  | 0.015163 | 50 × 50                | 1.010868  | 0.003961 |
| Reference $k_{eff} = 1.010909$ |           |          |                        |           |          |

**Table 3.**  
Neutron multiplication factor ( $k_{eff}$ ) convergence for homogeneous reactor 2D.

| Triangular elements |                        | Quadrilateral elements |                        |
|---------------------|------------------------|------------------------|------------------------|
| Number of elements  | PRE <sub>max</sub> (%) | Number of elements     | PRE <sub>max</sub> (%) |
| 05 × 05 × 02        | 1.311736               | 05 × 05                | 5.882691E-02           |
| 10 × 10 × 02        | 0.348277               | 10 × 10                | 5.882691E-02           |
| 15 × 15 × 02        | 0.160379               | 15 × 15                | 5.882691E-02           |
| 20 × 20 × 02        | 9.976203E-02           | 20 × 20                | 5.882691E-02           |
| 40 × 40 × 02        | 4.241142E-02           | 40 × 40                | 5.882691E-02           |
| 50 × 50 × 02        | 3.561607E-02           | 50 × 50                | 5.882691E-02           |

**Table 4.**  
Power convergence for homogeneous reactor 2D.

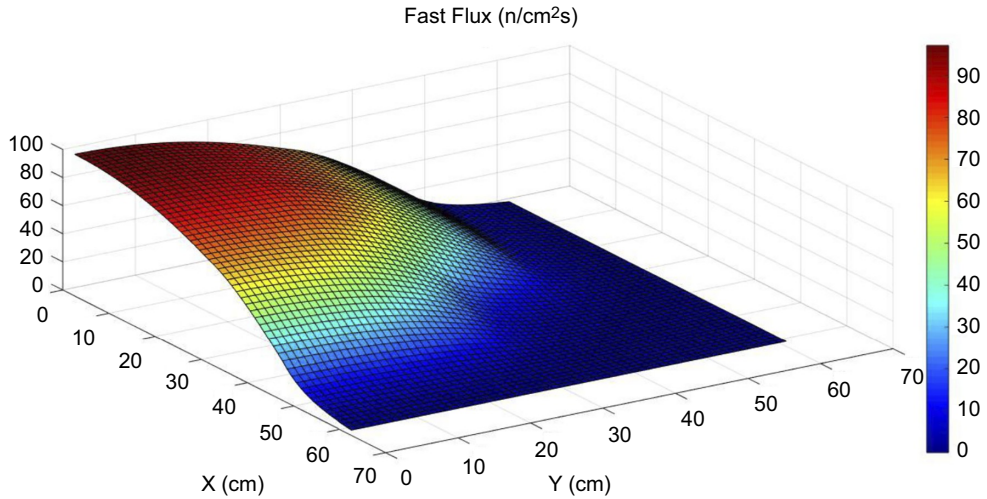
|                                                                     |             |             |             |             |             |
|---------------------------------------------------------------------|-------------|-------------|-------------|-------------|-------------|
| <div>Subtitle:<br/>Reference<br/>PRE-GFEM (T)<br/>PRE-GFEM(Q)</div> | 2.3866      | 2.1531      | 1.7089      | 1.0973      | 0.3781      |
|                                                                     | -0.3000E-01 | -0.1677E-01 | -0.1229E-02 | 0.1239E-01  | -0.1728E-01 |
|                                                                     | -0.2883E-01 | -0.2336E-01 | -0.1223E-01 | -0.1458E-02 | -0.2889E-01 |
|                                                                     |             | 1.9426      | 1.5418      | 0.9900      | 0.3412      |
|                                                                     |             | -0.3964E-02 | 0.4540E-02  | 0.1148E-01  | 0.1237E-01  |
|                                                                     |             | -0.9935E-02 | -0.6486E-03 | 0.9828E-02  | 0.3013E-01  |
|                                                                     |             |             | 1.2238      | 0.7858      | 0.2708      |
|                                                                     |             |             | 0.1651E-01  | 0.1537E-01  | -0.1248E-01 |
|                                                                     |             |             | 0.1740E-01  | 0.2656E-01  | 0.3829E-01  |
|                                                                     |             |             |             | 0.5046      | 0.1739      |
|                                                                     |             |             |             | 0.1147E-01  | -0.5026E-01 |
|                                                                     |             |             |             | 0.4348E-01  | 0.5883E-01  |
|                                                                     |             |             |             |             | 0.0599      |
|                                                                     |             |             |             |             | -0.3247E+00 |
|                                                                     |             |             |             |             | 0.2234E-01  |

T = triangular elements    Q = quadrilateral elements

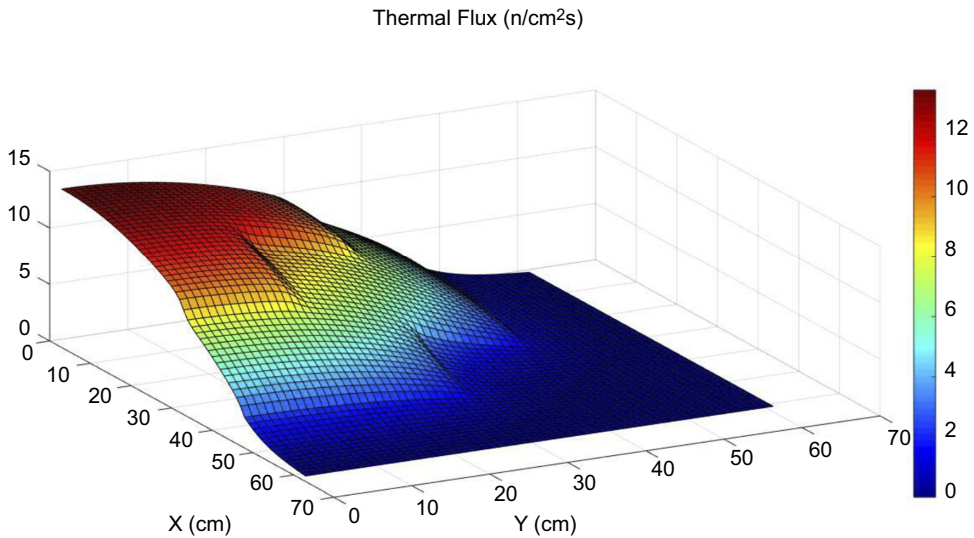
**Figure 10.**  
Normalized assembly power distribution of 1/4 of the homogeneous reactor 2D reactor core calculated using the GFEM with 50 × 50 × 02 triangular elements and 50 × 50 quadrilateral elements.

## 7. EPRI-9 2D reactor

**Figure 11** shows the neutron flux for  $g = 1$  (fast flux) and  $g = 2$  (thermal flux) neutrons for the EPRI-9 reactor. **Tables 5** and **6** show the convergence of the eigenvalue and power, respectively. **Figure 12** shows the values of the reference



(a)



(b)

**Figure 11.** EPRI-9 2D reactor (a) fast neutron flux and (b) thermal neutron flux obtained from the GFEM method with  $60 \times 60$  partitions.

| Triangular elements            |           |            | Quadrilateral elements |           |          |
|--------------------------------|-----------|------------|------------------------|-----------|----------|
| Number of elements             | $k_{eff}$ | PRE(%)     | Number of elements     | $k_{eff}$ | PRE(%)   |
| 04 × 04 × 02                   | 0.920963  | 0.67676574 | 04 × 04                | 0.922878  | 0.470277 |
| 08 × 08 × 02                   | 0.9249506 | 0.24679503 | 08 × 08                | 0.9256897 | 0.167083 |
| 16 × 16 × 02                   | 0.9265921 | 0.06975991 | 16 × 16                | 0.9268001 | 0.047323 |
| 32 × 32 × 02                   | 0.9270473 | 0.02067240 | 32 × 32                | 0.9271132 | 0.013562 |
| 40 × 40 × 02                   | 0.9271058 | 0.01436347 | 40 × 40                | 0.9271540 | 0.009163 |
| 60 × 60 × 02                   | 0.9271649 | 0.00798826 | 60 × 60                | 0.9271954 | 0.004701 |
| Reference $k_{eff} = 0.927239$ |           |            |                        |           |          |

**Table 5.**  
Neutron multiplication factor ( $k_{eff}$ ) convergence for EPRI-9 2D.

| Triangular elements |                        | Quadrilateral elements |                        |
|---------------------|------------------------|------------------------|------------------------|
| Number of elements  | PRE <sub>max</sub> (%) | Number of elements     | PRE <sub>max</sub> (%) |
| 04 × 04 × 02        | 6.431370               | 04 × 04                | 11.178396              |
| 08 × 08 × 02        | 3.8945167              | 08 × 08                | 3.901778               |
| 16 × 16 × 02        | 1.4236291              | 16 × 16                | 1.423513               |
| 32 × 32 × 02        | 0.5641198              | 32 × 32                | 0.549069               |
| 40 × 40 × 02        | 0.4408529              | 40 × 40                | 0.429145               |
| 60 × 60 × 02        | 0.3119380              | 60 × 60                | 0.305763               |

**Table 6.**  
Power convergence for EPRI-9 2D.

|                                                                                   |               |            |
|-----------------------------------------------------------------------------------|---------------|------------|
| 1.4458                                                                            | 1.2188        | 0.8490     |
| -0.2773E+00                                                                       | -0.6399E-01 - | 0.1459E+00 |
| -0.2504E+00                                                                       | -0.5726E-01   | 0.1254E+00 |
| <b>Subtitle:</b><br><b>Reference</b><br><b>PRE-GFEM (T)</b><br><b>PRE-GFEM(Q)</b> |               | 0.6072     |
|                                                                                   |               | 0.3119E+00 |
|                                                                                   |               | 0.3058E+00 |

T = triangular elements  
Q = quadrilateral elements

**Figure 12.**  
Normalized assembly power distribution of ¼ of the EPRI-9 2D reactor core calculated using the GFEM with 60 × 60 × 02 triangular elements and 60 × 60 quadrilateral elements.

power obtained with the FDM and the percentage relative errors calculated through the comparison made with the power obtained with the GFEM using  $60 \times 60$  partitions.

For the EPRI-9 problem, the bilinear rectangular elements presented better convergence in the analysis of the eigenvalue  $k_{eff}$ . However, when dealing with power, the bilinear triangular elements present results closer to the reference for cases with up to  $8 \times 8$  partitions. For the bilinear triangular elements, the smallest error presented for the eigenvalue and power are 0.007988 and 0.311938, respectively. In turn, considering the bilinear rectangular elements, the smallest error presented for the eigenvalue and power are 0.004701 and 0.305763, respectively.

## 8. Conclusions

In order to solve the NDE in steady state, this chapter presents a numerical strategy based on GFEM using linear form functions for both triangular and quadrilateral elements. The results presented allow us to conclude that:

- Despite presenting a more complex mathematical formulation than FDM, GFEM can be adapted with relative ease to solve the NDE.
- Triangular elements, which are elements that have greater versatility in representing irregular geometries and generating unstructured meshes, can produce results as satisfactory as rectangular elements.

When it comes to numerically solving linear PDEs, FEM, despite having the ability to handle different types of geometry and presenting good convergence rates, also has the disadvantage of requiring greater computational effort compared to FDM, for example, both due to the amount of memory required for its implementation and execution and the calculation time, which is influenced by the characteristics of the large-scale linear system intrinsic to its mathematical formulation. Over the years, various research efforts have been conducted aiming to overcome these disadvantages. In the area of Reactor Physics, for example, there are power reconstruction techniques [11, 18] that treat the neutron diffusion problem as a Dirichlet problem, where the fluxes over the faces of the FAs are prescribed values generated through polynomial expansions based on solutions obtained with the NEM, a method that is widely used in Reactor Physics. Some research efforts seek to reduce computational effort by simplifying the mathematical formulation of FEM, ensuring that there is no loss of stability and/or accuracy [19]. Other research focuses on accelerating FEM calculation time through more modern computational resources, such as cloud computational parallelization [12]. As it is actually a family of methods, FEM remains a growing area of research in numerical methods, particularly in nuclear engineering and several other fields of engineering.

## **Author details**

Lenilson Moreira Araujo\* and Anderson Alvarenga de Moura Meneses  
Federal University of Western Pará, Institute of Engineering and Geosciences,  
Laboratory of Computational Intelligence, Santarém, PA, Brazil

\*Address correspondence to: lenilson.moreira1@gmail.com

## **IntechOpen**

---

© 2025 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Kok KD. Nuclear Engineering Handbook. New York: CRC Press Taylor & Francis Group; 2009
- [2] Duderstadt JJ, Hamilton LJ. Nuclear Reactor Analysis. EUA: John Wiley and Sons, Inc; 1976
- [3] Lewis EE, Miller Jr WF. Computational Methods of Neutron Transport. New York: Jhon Wiley & Sons; 1985
- [4] Pessoa PO, Silva FC, Martinez AS. Finite difference applied to the reconstruction method of the nuclear power density distribution. *Annals of Nuclear Energy*. 2016;**92**:378–390
- [5] Finnemann H, Bennewitz F, Wagner MR. Interface current techniques for multidimensional reactor calculations. *Atomkernenergie*. 1977;**30**:123
- [6] Wang Y, Chi H, Ma Y. Nodal expansion method based reduced-order model for control rod movement. *Annals of Nuclear Energy*. 2024;**198**:110279
- [7] Smith KS, 1979. An Analytic Nodal Method for Solving the Two-group, multidimensional, Static and Transient Neutron Diffusion Equations. Massachusetts Institute of Technology. Ph.D. Thesis
- [8] Silva A, Pessoa PO, Silva FC, Martinez AS. Two-dimensional analytical solution for nodal calculation of nuclear reactors. *Annals of Nuclear Energy*. 2017;**100**:76–81
- [9] Raj M, Ahmed N, Singh S. Analytical nodal method for solution of neutron diffusion equation in polar coordinates. *Annals of Nuclear Energy*. 2022;**165**:108659
- [10] Hosseini SA, Derakhshandeh FS. Galerkin and generalized least squares finite element: A comparative study for multi-group diffusion solvers. *Progress in Nuclear Energy*. 2015;**85**:473–490
- [11] Pessoa PIO, Araujo LM, Silva FC. A strategy for pin power reconstruction based on classic Galerkin variational formulation. *Progress in Nuclear Energy*. 2017;**104**:251–263
- [12] de Moura Meneses AA, Araujo LM, Schirru R. A GPU-accelerated linear system solution for the Galerkin finite element method applied to neutron diffusion equation. *Nuclear Engineering and Design*. 2024;**421**:113103
- [13] Trainor J, Eaton MD, Kópházi J, Wilson SG, Latimer C, Smith L, Baker D, Jordan I. NURBS-enhanced finite element spatial discretisation methods for the steady-state multigroup neutron diffusion equation. *Annals of Nuclear Energy*. 2025;**223**:2025
- [14] Adams RA. Sobolev Spaces, Academic Press Inc. 1975
- [15] Johnson C. Numerical Solution of Partial Differential Equation by the Finite Element Method. New York: *Cambridge University Press*; 1990
- [16] Burden RL, Faires JD. Numerical Analyses, Ninth Edition. Boston: Brooks/Cole, Cengage Learning; 2010
- [17] Aboanber AE, Hamada YM. Generalized Runge–Kutta method for o- and three-dimensional space–time

diffusion equations with a variable time step. Ann. Annals of Nuclear Energy. 2008;**35**:1024–1040

[18] Pessoa PIO, Araujo LM, Silva FC, Menezes AAM. Numerical methods applied to pin power reconstruction based on coarse-mesh nodal calculation. ANNALS OF NUCLEAR ENERGY. 2018;**118**:291–312

[19] Lenilson MA, Do Carmo EGD, da Silva FC. Galerkin partial least-square SN (GpLS - SN) method for fixed source problems in the neutrons transport theory. Annals of Nuclear Energy. 2016;**90**:175–190

# Numerical Analysis of the Enriched Mixed Method, Evaluating Computational Costs and Errors through Comparisons with Formulations in Continuous Functions, Discontinuous Functions, and Mixed Finite Elements

*Jorge L. Díaz Calle, Roger Mestas Chávez and  
Aldo Sarmiento Guevara*

## Abstract

In diffusion problems, mixed finite element formulation are implemented to simultaneously approximate primal and dual variables. A new discrete mixed approximation has been devised by Devloo, which utilizes an enhanced version of conventional Poisson-compatible spaces. The objective of this chapter was to assess the computational cost of implementing this new formulation in conjunction with three other classical formulations to solve three Poisson problems that have known exact solutions in two and three dimensions. The goal is not to determine the preferable formulation, but rather to stress the computational expense and the errors that are derived for both the primal and dual quantities. Weak formulations include the mixed method, the enriched mixed method, and the finite element method, which uses both continuous and discontinuous functions. The computational expenses of the assembly and solution procedures are evaluated, and the errors in  $L^2$  and  $H^1$  norms are computed. It is evident that the computational expenditure of the enriched version is lower than that of the non-enriched version when these costs are compared to the resulting errors. However, their approximation errors are quite similar. The improved version is perceived as an optimized iteration of the mixed finite element method. Three of these formulations exhibit a close equivalence.

**Keywords:** enriched mixed method, mixed method, classical FEM, discontinuous Galerkin, computational cost

## 1. Introduction

The authors are presently engaged in research pertaining to the agro-industry in the Arequipa region of Peru. The mathematical models that facilitate our comprehension of the behavior of thermal and chemical processes of interest examine the transport equation in its transient form. Our analyses revealed the necessity for adaptability in the advanced finite element techniques we have created. The algorithms utilize the traditional finite element approach, namely its discontinuous Galerkin variation, as well as the mixed method and its enriched variant. A persistent inquiry arises: what is the computational expense associated with these finite element numerical methods, and what is the error ratio, particularly in comparison to the enriched variation of the mixed method? This chapter has been constructed to analyze the Poisson equation across three distinct problems: one smooth, one exhibiting a strong gradient, and a third that is highly oscillatory, all possessing known precise solutions. The objective is not to identify the most efficient formulation but to evaluate the computational expense and the mistakes associated with both the primal and dual variables while employing any of the examined techniques.

Finite element simulations produce approximate solutions by subdividing the domain into a mesh consisting of a finite number of elements. Numerous finite element formulations have been developed employing both continuous and discontinuous finite element approximation spaces. Two prior assessments were given at the XLI Cilamce Congress in 2020 [1, 2]. The  $H^1$ -conforming weak formulation has demonstrated efficacy in approximating the primal variable, and its computational cost can be optimized through static condensation. An alternative viable option is to employ approaches primarily reliant on discontinuous approximation spaces, specifically referred to as discontinuous Galerkin spaces ( $DG$ ). These spaces comprise discontinuous piecewise polynomials. The utilization of these places offers advantages, including the ability to adopt  $h$ ,  $p$ , and  $hp$ -adaptive techniques, as well as a local conservation property. A practical drawback is that the number of degrees of freedom (DoFs) is significantly more for a given mesh compared to classical conforming Galerkin algorithms, and static condensation is ineffective.

When the accuracy of the dual variable (flux) is paramount, approximation using the gradient of  $H^1$  or  $DG$  solutions leads to reduced precision. An alternative formulation is the dual mixed  $H(\text{div})$ -conforming approach, as cited in the studies by Brezzi and Fortin [3], Raviart and Thomas [4], and Nedelec [5]; it depends on simultaneous approximations of both the primal and dual variables. The approximation spaces for the primal and dual variables must be compatible and satisfy the equilibrium requirement to ensure a consistent and stable numerical method. Devloo has proposed to augment the dual variable space by increasing the polynomial order of the vector bubble polynomials in the finite elements, employing the  $H(\text{div})$ -conforming formulation [6, 7]. To sustain the equilibrium condition, utilizing balanced spaces, Devloo must also elevate the order of the polynomials for the fundamental variable space.

This study aims to mathematically examine Devloo's new suggestion, an enriched mixed formulation. A preliminary research conducted in the study by Calle et al. [8] did not account for the formulation including discontinuous functions. Augmented by one degree the polynomials employed in the formulation of the vector bubble

functions within the elements. Our analysis confirms the theoretical convergence rate for smooth solutions, and we present a graph illustrating exponential convergence as the interpolation order varies from  $p = 2$  to  $p = 7$ , against the number of condensed DoFs required (CDoFs), utilizing three fixed uniform refinement levels. We also confirm that the quantity of condensed DoFs corresponds to that of the mixed formulation without enrichment.

Comparing the mistakes from the mixed and the enriched mixed methods reveals that the enriched method of order  $p$  corresponds with the mixed method of order  $(p + 1)$ . Therefore, we perform a direct analysis to determine the levels of the methods necessitating comparison. The improved version has fewer DoFs, leading to decreased computing costs while attaining virtually same error rates. Driven by these attributes, we opt to examine the processing time, assembly technique, and resolution of the linear system concurrently with three alternative formulations: the mixed non-enrichment, the classical  $H^1$  formulation (Conforming  $H^1$ ), and a discontinuous Galerkin (DG) formulation. The computing cost of the assembly and solver procedures was quantified in clock ticks.

The implemented formulations are: (1) the classical  $H^1$ -conforming weak formulation (continuous); (2) the non-symmetric discontinuous Galerkin formulation by Baumann, [9]; (3) a mixed  $H(\text{div})$ -conforming formulation [6]; (4) an enriched mixed  $H(\text{div})$ -conforming formulation [7, 10]. Static condensation is consistently employed for continuous, mixed  $H(\text{div})$ -conforming and enhanced mixed  $H(\text{div})$ -conforming formulations. For further information on the application of static condensation in continuous and mixed  $H(\text{div})$ -conforming formulations, see the study by Forti et al. [11], and for the enriched mixed  $H(\text{div})$ -conforming formulation, consult the study by Farias et al. [10].

All formulations were executed utilizing the object-oriented scientific computational environment NeoPZ, with development facilitated by the extensive resources offered by NeoPZ [12]. NeoPZ is a versatile scientific computing framework designed for the development of finite element algorithms, incorporating diverse element geometries, variational formulations, and approximation spaces. In three dimensions, it encompasses hexahedral, tetrahedral, prism, and pyramid elements [13]. The framework comprises modules for several technologies, encompassing system resolution, geometric approximation of finite elements, and finite element approximation spaces. It facilitates the execution of localized mesh refinements and the determination of the approximation order. To substantiate the concepts presented in this research, numerical experiments were conducted on a MacBook Pro, including a six-core Intel Core i7 CPU (2.2 GHz) and 32 GB of RAM (DDR4 2,400 MHz).

## 2. Finite element formulations

Examine the Poisson model defined within a polygonal domain  $\Omega$

$$-\nabla \cdot (\mathbf{D}\nabla u) = f, \text{ in } \Omega \quad (1)$$

$$u = 0, \text{ on } \partial\Omega, \quad (2)$$

where  $f \in L^2(\Omega)$ .  $\mathbf{D}$  is a symmetric and uniformly positive definite matrix.  $\partial\Omega$ , denotes the Lipschitz boundary of the  $\Omega$ . The domain  $\Omega$  may exist in either  $\Omega \subset \mathbb{R}^2$  or  $\Omega \subset \mathbb{R}^3$

Let  $M_h$  denote a mesh of finite elements on  $\Omega$ , with each finite element denoted by  $\Omega_E$ . A master element  $\hat{\Omega}$  is linked to the elements  $\Omega_E$  using a geometric diffeomorphism  $F_{\Omega_E} : \hat{\Omega} \rightarrow \Omega_E$ .

In two-dimensional (2D) scenarios,  $\hat{\Omega}$  is a square defined as  $[-1, 1]^2$ , and in three-dimensional (3D) cases, it is a cube defined as  $[-1, 1]^3$ .

$U_p(\hat{\Omega})$  constitutes a collection of hierarchical polynomial functions of degree  $p$  defined on  $\hat{\Omega}$ . Induced by the diffeomorphism  $F_{\Omega_E}$ , a conventional mapping of scalar functions  $\mathbf{F}_{\Omega_E}$  is employed.

Subsequently,

$$U_p(\Omega_E) = \mathbf{F}_{\Omega_E}(U_p(\hat{\Omega})) = \left\{ \varphi \mid \varphi \circ \mathbf{F}_{\Omega_E} \in U_p(\hat{\Omega}) \right\} \quad (3)$$

is employed to establish the approximation space for the pressure in the four formulations presented below. For details about the construction of the aforementioned form functions, refer to the study by Devloo et al. [14].

## 2.1 Conforming $H^1$ formulation

The following is the conventional weak formulation for the Poisson model given by Eqs. (1) and (2):

Determine  $u_h \in H_0^1(\Omega)$

$$(\mathbf{D}\nabla u_h, \nabla v_h)_\Omega = (f, v_h), \forall v_h \in H_0^1(\Omega). \quad (4)$$

The space  $H_0^1(\Omega)$  is made up of all functions  $v_h$  defined on a domain  $\Omega$  for which both  $v_h$  and its first weak derivative are square-integrable over  $(\Omega)$ , and have zero value on the boundary  $\partial\Omega$ .

## 2.2 Discontinuous Galerkin formulation

The discontinuous Galerkin weak formulation is developed within the framework of broken polynomial spaces.

$$V_p(M_h) = \{v \in L^2(\Omega); v|_{\Omega_E} \in U_p(\Omega_E), \forall \Omega_E \in M_h\}. \quad (5)$$

The discrete weak formulation defined in the study by Baumann [9] for the Poisson model given by Eqs. (1) and (2) is:

Determine  $u_h \in V_p(M_h)$  such that

$$\begin{aligned} & \sum_{E=1}^{N_E} \int_{\Omega_E} \nabla u_h \cdot \nabla v_h d\Omega_E + \int_{\partial\Omega} (u_h \nabla v_h \cdot \mathbf{n} - v_h \nabla u_h \cdot \mathbf{n}) ds + \\ & + \int_{\Gamma} (\langle \nabla v_h \cdot \mathbf{n}_- \rangle [u_h] - \langle \nabla u_h \cdot \mathbf{n}_- \rangle [v_h]) ds = \sum_{E=1}^{N_E} \int_{\Omega_E} f v_h d\Omega_E \end{aligned} \quad (6)$$

for each  $v_h \in V_p(M_h)$ .

$\Gamma$  represents the amalgamation of all interelement boundaries referred to as “interfaces.” In the 2D scenario, each interface  $\Gamma_{rl} \subset \Gamma$  serves as a common boundary between two quadrilateral elements designated as  $\Omega_{E_r}$  (right element) and  $\Omega_{E_l}$  (left element).

The operators include

$$\langle \nabla v_h \cdot \mathbf{n}_- \rangle = \frac{1}{2} (\nabla v_{E_r} \cdot \mathbf{n}_{E_r} + \nabla v_{E_l} \cdot \mathbf{n}_{E_l}) \quad (7)$$

and

$$[v_h] = \left( v_h|_{\partial\Omega_{E_r} \cap \Gamma_H} - v_h|_{\partial\Omega_{E_l} \cap \Gamma_H} \right) \quad (8)$$

and  $\mathbf{n}_E$  refers to the outward unit normal of the boundary  $\partial\Omega_E$ . Also,  $\mathbf{n}$  refers to the outward unit normal of the boundary  $\partial\Omega$ .

This formulation will not be applicable to a 3D scenario.

### 2.3 H(div)-conforming mixed finite element formulation

A few of balanced spaces are required for a mixed formulation. Various classic families of suitable spatial layouts exist in the literature. This research utilizes the spaces constructed by de Siqueira et al. [6].

Considering finite-dimensional approximation spaces, a vector function space  $\mathbf{V}_p \subset \mathbf{H}(\text{div}, \Omega)$ , and a scalar function space  $U_p \subset L^2(\Omega)$ , we examine the discrete mixed formulation for the Poisson model given by Eqs. (1) and (2):

Determine  $\sigma_h \in \mathbf{V}_p \subset \mathbf{H}(\text{div}, \Omega)$ ,  $u_h \in U_p \subset L^2(\Omega)$  satisfying

$$(\sigma_h, \mathbf{q})_\Omega - (u_h, \nabla \cdot \mathbf{q})_\Omega = 0, \forall \mathbf{q} \in \mathbf{V}_p, \quad (9)$$

$$-(\nabla \cdot \sigma_h, v)_\Omega + (f, v)_\Omega = 0, \forall v \in U_p. \quad (10)$$

To establish  $U_p$  and  $\mathbf{V}_p$ , scalar and vector and polynomial approximation spaces  $\hat{U}_p$  and  $\hat{\mathbf{V}}_p$  are established within the master element  $\hat{\Omega}$ , which are presumed to be divergence compatible:  $\nabla \cdot \hat{\mathbf{V}}_p = \hat{U}_p$ .

Local approximation spaces  $\mathbf{V}_p(\Omega_E)$  and  $U_p(\Omega_E)$  are defined on the computational elements  $\Omega_E$ , tracing back to the polynomial spaces  $\mathbf{V}_p(\hat{\Omega})$  and  $U_p(\hat{\Omega})$ . The Piola transformation  $\mathbf{F}_{\Omega_E}^{div}$  is used.

$$\mathbf{V}_p(\Omega_E) = \mathbf{F}_{\Omega_E}^{div} \left( \mathbf{V}_p(\hat{\Omega}) \right) = \{ \mathbf{q} | \hat{\mathbf{q}} = |J_{F_{\Omega_E}}|^{-1} \mathbf{q} \circ \mathbf{F}_{\Omega_E} \in \mathbf{V}_p(\hat{\Omega}) \},$$

where  $J_{F_{\Omega_E}}$  denotes the Jacobian of  $F_{\Omega_E}$ , and  $|J_{F_{\Omega_E}}| = \det(J_{F_{\Omega_E}})$  is presumed to be positive.

In a straightforward scenario, to construct  $\mathbf{V}_p$  and  $U_p$ , the polynomial order in  $\hat{\mathbf{V}}_p$  is  $(p+1)$ , whereas in  $\hat{U}_p$  it is  $p$ , to satisfy the *inf-sup* requirement. For further information, refer to the study by de Siqueira [6].

### 2.4 Enriched mixed finite element formulation

The case of the enriched formulation proposed by Devloo is described in somewhat greater detail. For further information, refer to the studies by Castro et al. [15], Devloo et al. [16], and Farias et al. [10].

Examine two divergence-compatible spaces: a vector polynomial space referred to as  $\hat{\mathbf{V}}_p$  and a scalar polynomial space designated as  $\hat{U}_p$ , both delineated in  $\hat{\Omega}$  (master element) as described in Section 2.3.

Let us consider that a direct decomposition

$$\hat{\mathbf{V}}_p = \hat{\mathbf{V}}_p^\partial \oplus \hat{\mathbf{V}}_p^\circ \quad (11)$$

is acceptable, where  $\hat{\mathbf{V}}_p^\circ$  represents the flux functions with zero normal components on  $\partial\hat{\Omega}$ . In contrast, the vector functions in  $\hat{\mathbf{V}}_p^\partial$  are anticipated to display normal components of degree  $p$  on  $\partial\hat{\Omega}$ .

In this context, an augmented version  $\hat{\mathbf{V}}_p^+$  is created by integrating higher-degree internal form functions from the initial vector space at level  $(p + 1)$ , while preserving the original boundary fluxes at level  $p$

$$\hat{\mathbf{V}}_p^+ = \hat{\mathbf{V}}_p^\partial \oplus \hat{\mathbf{V}}_{p+1}^\circ. \quad (12)$$

To define the pair of approximation spaces, at least four important aspects must be taken into account; see Section 3 in the study by Castro et al. [15].

Various configurations of mixed formulations can be developed, utilizing distinct approximation spaces  $\hat{\mathbf{V}}_p$  and  $\hat{U}_p$ , based on regular meshes to achieve different accuracy orders for the primal variable, while maintaining a constant accuracy order for the dual variable. The elements may consist of quadrilaterals, triangles, tetrahedra, hexahedra, or prismatic geometries. In the original upgraded version, Devloo elevates the degree of the internal shape functions to level  $(p + n)$ , where  $n \in \mathbb{N}$ . Nevertheless, we found it very appealing to practically utilize the scenario where  $n = 1$ .

The associated enriched scalar spaces (potential) are now  $\hat{U}_p^+ = \nabla \cdot \hat{\mathbf{V}}_p^+ = \hat{U}_{p+1}$ .

The two approximation spaces  $\mathbf{V}_p \subset \mathbf{H}(\text{div}, \Omega)$  and  $U_p \subset L^2(\Omega)$  are delineated based on local spatial configurations  $\{\mathbf{V}_p^+(\Omega_E), U_p^+(\Omega_E)\}$ , derived from  $\{\hat{\mathbf{V}}_p^+, \hat{U}_p^+\}$  using Piola transformation. See Subsection 2.4. Transformations as shown in the study by Devloo et al. [16].

### 3. Numerical experiments using Poisson model

In the next subsections, we analyze the Poisson model with a source function  $f$  established to identify three distinct exact solutions. In the first scenario, we take into consideration an exact solution with sinusoidal behavior that is suitably smooth. The second solution possesses a exact solution characterized by a very high gradient. In the third scenario, the model possesses an exact solution characterized by pronounced oscillatory behavior. The second scenario considers one section with a 2D domain and another section with a 3D domain.

In the first subsection, when we examined the Poisson model with a smooth solution, we employed four algorithms: the  $H^1$ -conforming, the Discontinuous Galerkin, the mixed, and the enriched mixed methods. The aim is to ascertain the convergence rate of the enriched mixed algorithm and to verify the convergence rates of other algorithms documented in the scientific literature. The results are presented for the polynomial order  $p = 2$  across three progressively evenly uniformly refined meshes.

The second subsection examines the Poisson model with a solution with a strong gradient in the 2D scenario. In this instance, when the errors are consistently substantial, a comparative analysis of the errors from the four methods in the  $\|\cdot\|_{L^2}$

and  $\|\cdot\|_{H^1}$  norms will be shown. A sufficiently refined mesh is employed, and the polynomial degree is changed from  $p = 2$  to  $p = 7$ . The section finishes by presenting the computational expenses associated with the  $H^1$ -conforming, discontinuous Galerkin, and enriched mixed algorithms in comparison to those of the mixed algorithm.

Owing to the attributes of the discontinuous Galerkin algorithm, the second section demonstrates that its computational expense is significantly greater than that of the other three algorithms referenced; hence, this algorithm will be excluded from subsequent comparisons.

The third section examines the same Poisson model, characterized by a solution with a high slope in a 3D domain. Only the subsequent algorithms are employed:  $H^1$ -conforming, mixed, and enhanced enriched mixed. Errors are evaluated using the  $\|\cdot\|_{L^2}$  and  $\|\cdot\|_{H^1}$  norms. A coarser mesh is employed to prevent an excessive quantity of DoFs. The polynomial degree ranges from  $p = 1$  to  $p = 5$ , hence decreasing processing time. The comparative computing expense of the three aforementioned techniques is also presented.

In the final section, we examine the Poisson model featuring a strongly oscillatory solution in two dimensions. This enables us to evaluate the efficacy and stability of the implementations within the NeoPZ computational framework. We persist in utilizing solely the three algorithms, excluding the Discontinuous Galerkin technique. A sufficiently fine mesh and polynomial orders ranging from  $p = 2$  to  $p = 7$  will be employed once more.

### 3.1 Two-dimensional Poisson model exhibiting a sinusoidal exact solution

Examine the Poisson model given by Eqs. (1) and (2) with the subsequent exact solution:

$$u_{exact}(x,y) = \sin(\pi x) \sin(\pi y), \quad (13)$$

The source function  $f \in L^2(\Omega)$  corresponds to  $f = -\Delta u_{exact}$  in the domain  $\Omega = [-1, 1]^2$ , according to Dirichlet boundary conditions. The exact solutions for pressure and flow are illustrated in **Figures 1** and **2**, respectively.

In the context of the mixed approach, it is presumed that a direct decomposition  $\hat{\mathbf{V}}_p = \hat{\mathbf{V}}_p^\partial \oplus \hat{\mathbf{V}}_p^\circ$  is valid, where  $\hat{\mathbf{V}}_p^\circ$  denotes the flux functions with zero normal components on  $\partial\hat{\Omega}$ . Conversely, the functions in the existing model are utilized to verify the characteristics of the improved mixed method. The initial verification conducted pertains to the convergence behavior of the approach as the order increases from  $p = 2$  to  $p = 7$ .

The theory illustrates convergence in **Figure 3**.

The mixed finite element method of order  $p$  and the enriched mixed finite element method of order  $p$  were confirmed to possess an equivalent quantity of condensed DoFs. The convergence rate of the mixed method exhibits a rate of  $p$ , whereas the enriched mixed method is determined to be  $(p + 1)$ , as illustrated in **Table 1**.

From this observation, we deduce that the comparison between the enriched mixed method of order  $p$  and the mixed method of order  $p$  is inequitable. The enhanced mixed method utilizes order  $(p + 1)$  for constructing  $\hat{U}_p^+$  and  $\hat{\mathbf{V}}_p^\partial$ , while it employs order  $(p + 2)$  for the building of  $\hat{\mathbf{V}}_{p+1}^\circ$ .

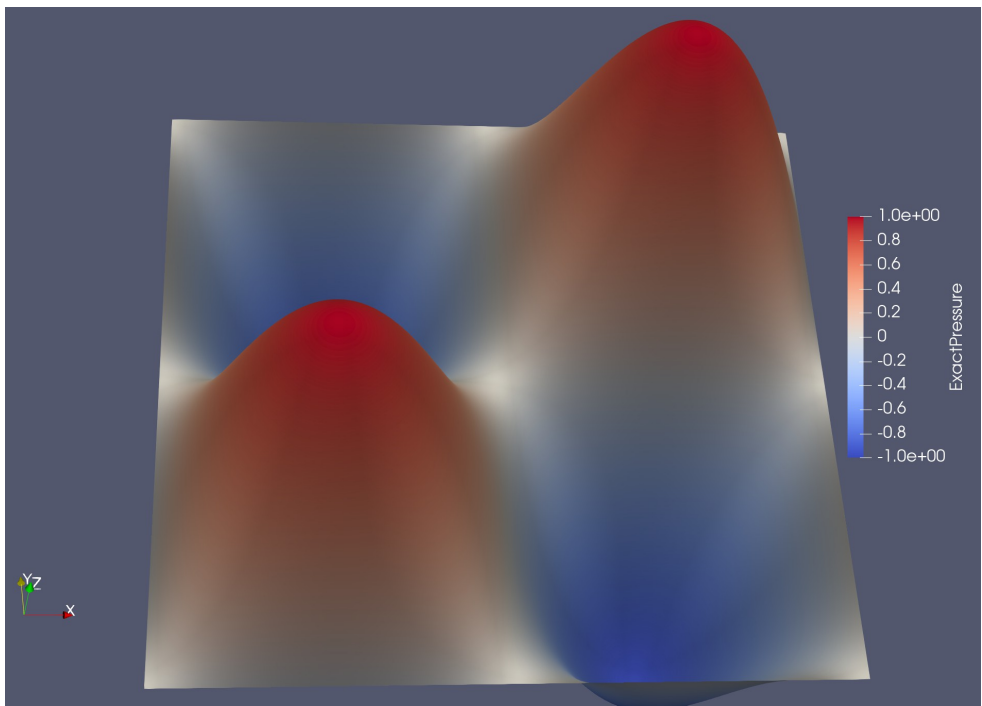


**Figure 1.**  
*The exact solution given by Eq. (13).*

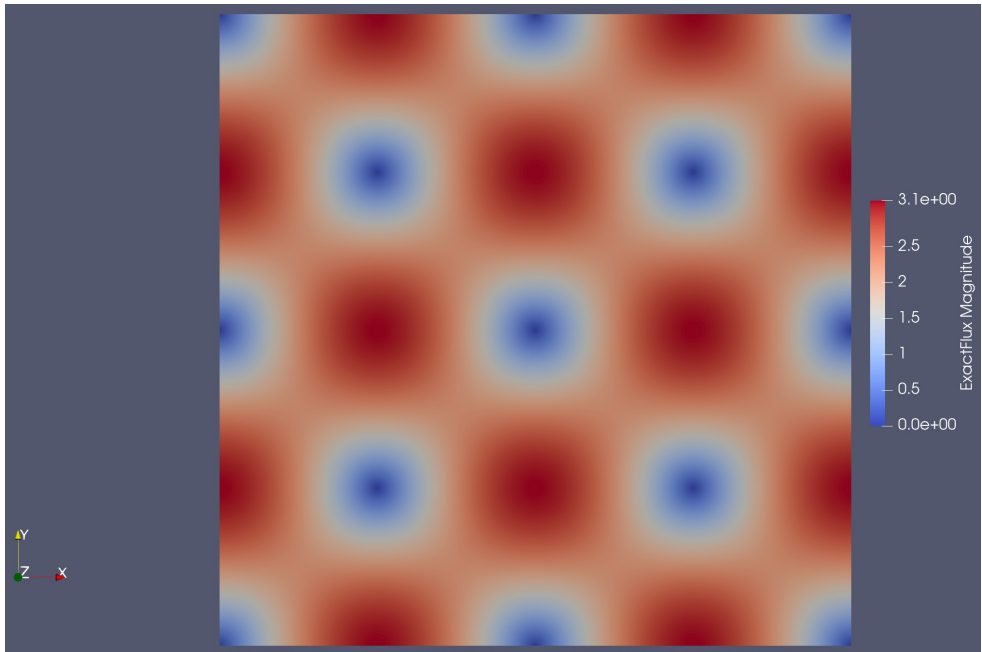


**Figure 2.**  
*The exact flux of the solution given by Eq. (13).*

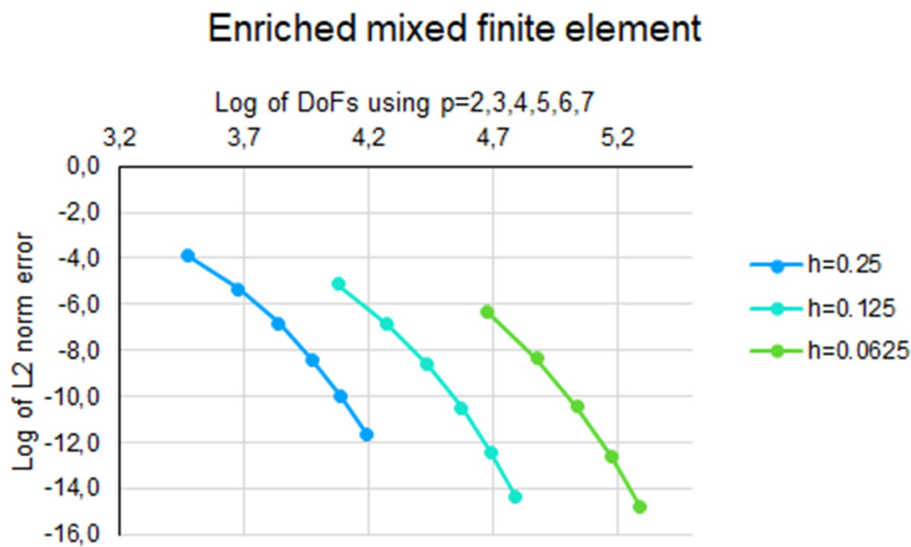


**Figure 3.**  
*Log(Error) versus Log(DoFs) for enriched mixed method, using three uniform  $h$ -refinements, varying  $p = 2$  until  $p = 7$ .*

| h                 | Enriched mixed |            | Mixed $H(\text{div})$ |            | Conforming $H^1$ |            | Discontinuous Galerkin |            |
|-------------------|----------------|------------|-----------------------|------------|------------------|------------|------------------------|------------|
|                   | Error          | Conv. rate | Error                 | Conv. rate | Error            | Conv. rate | Error                  | Conv. rate |
| $(\frac{1}{2})^2$ | 1.249<br>E-04  |            | 2.138<br>E-03         |            | 3.858<br>E-03    |            | 7.536<br>E-02          |            |
| $(\frac{1}{2})^3$ | 7.887<br>E-06  | 3.9851     | 2.688<br>E-04         | 2.9916     | 4.901<br>E-04    | 2.9796     | 2.346<br>E-02          | 1.6835     |
| $(\frac{1}{2})^4$ | 4.951<br>E-07  | 3.9936     | 3.368<br>E-05         | 2.9965     | 6.148<br>E-05    | 2.9948     | 6.268<br>E-03          | 1.9041     |
| $(\frac{1}{2})^5$ | 3.095<br>E-08  | 3.9997     | 4.213<br>E-06         | 2.9989     | 7.688<br>E-06    | 2.9994     | 1.591<br>E-08          | 1.9781     |

**Table 1.**  
*Verification of the convergence rate (Conv. rate) of the error in the  $\|\cdot\|_{L^2}$  (error) using  $p=2$ .*

This indicates that the corresponding version of the enhanced enriched mixed method of order  $p$  is the mixed method of order  $(p + 1)$ . Consequently, we opt to compare the mixed technique of order  $(p + 1)$  with the conventional FEM ( $H^1$ -conforming) of order  $(p + 2)$ , as well as the discontinuous Galerkin method, for a model that is not excessively smooth yet lacks any discontinuities. In the next sections, we analyze a model exhibiting pronounced oscillatory behavior and then a model distinguished by a strong gradient.

### 3.2 Two-dimensional Poisson model exhibiting a strong gradient in the exact solution

Examine the Poisson model given by Eqs. (1) and (2) with the subsequent exact solution:

$$u_{exact}(x,y) = \left(\frac{2}{5}\right) \left[ \frac{\pi}{2} + \arctan[10 - 200(x^2 + y^2)] \right], \quad (14)$$

The function  $f$  corresponds to  $f = -\Delta u_{exact}$  in the domain  $\Omega = [-1, 1]^2$ , according to Dirichlet boundary conditions. High gradients encircling the origin, with a radius of  $\frac{1}{4}$ . The exact solutions for pressure and flow are illustrated in **Figures 4** and **5**, respectively.

**Table 2** presents the errors using the  $\|\cdot\|_{L^2}$  for the four formulations of the strong gradient model. Compare the values of proximity in relation to the relevant sequence for the enriched mixed, mixed, and traditional FEM approaches. The disparities between mixed method and enriched mixed method are nearly comparable, differing by just millionths. The discontinuous Galerkin formulation demonstrates mistakes that are somewhat different from those of alternative approaches and involves diminished approximation.

**Table 3** presents the errors in the  $\|\cdot\|_{H^1}$  norm for the strong gradient model. The approximation of the mixed approaches about the  $\|\cdot\|_{H^1}$  is highly accurate. The mixed-method approach is the best. The traditional FEM method is now more precise than the mixed methods, while the discontinuous Galerkin method exhibits the lowest accuracy.

**Table 4** compares the computational cost (in clock ticks) of the enhanced mixed techniques, Discontinuous Galerkin, and conventional FEM to the time utilized by the mixed approach. The enhanced mixed methodology was observed to be up to 69% smaller than the mixed method, yet the resulting errors were significantly similar (see **Table 3**). The discontinuous Galerkin method exhibits the lowest effectiveness, attributed to the inadequacy of static condensation.

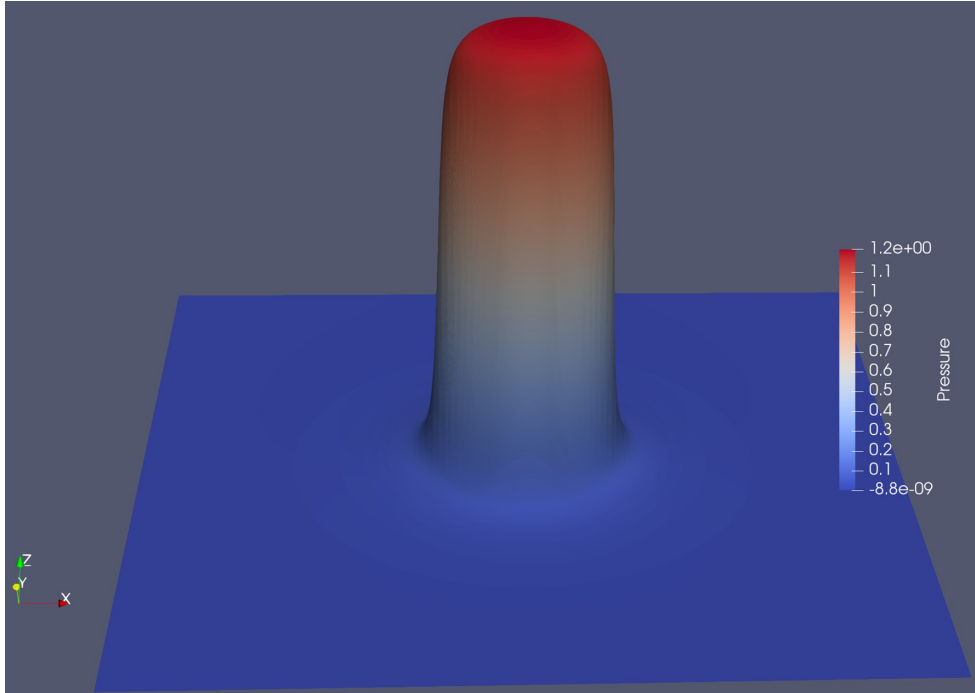


**Figure 4.**  
The exact solution given by Eq. (14) with a strong gradient.

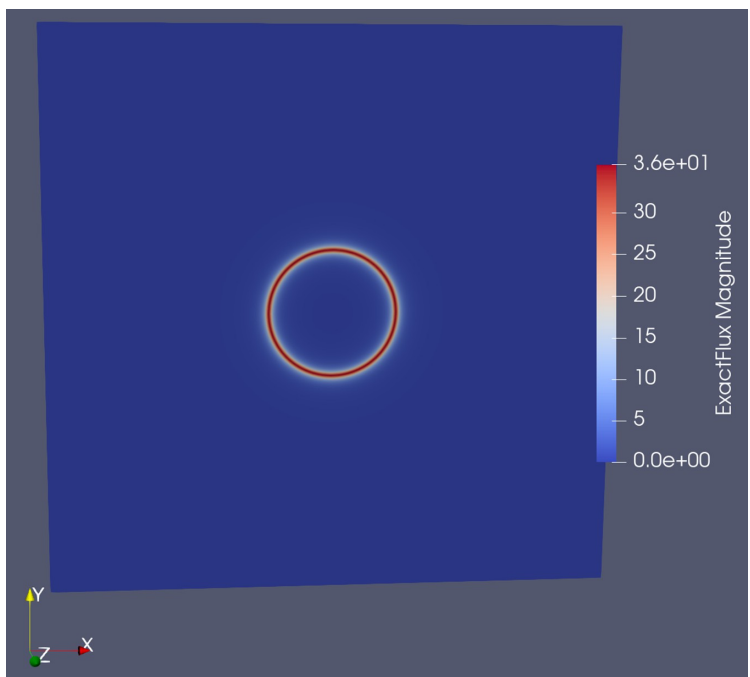


**Figure 5.**  
The exact flux of the solution given by Eq. (14).

| $p$ | Enriched mixed ( $p$ ) | Mixed ( $p + 1$ ) | $H^1$ -conforming ( $p + 2$ ) | DG ( $p + 2$ ) |
|-----|------------------------|-------------------|-------------------------------|----------------|
| 2   | 8.8852 E-01            | 8.8852 E-01       | 8.8853 E-01                   | 8.9694 E-01    |
| 3   | 3.7536 E-01            | 3.7536 E-01       | 3.7535 E-01                   | 3.7855 E-01    |
| 4   | 2.6311 E-01            | 2.6311 E-01       | 2.6311 E-01                   | 2.6138 E-01    |
| 5   | 1.1421 E-01            | 1.1421 E-01       | 1.1420 E-01                   | 1.1349 E-01    |
| 6   | 5.3405 E-02            | 5.3405 E-02       | 5.3403 E-02                   | 5.3905 E-02    |
| 7   | 2.1689 E-02            | 2.1689 E-02       | 2.1687 E-02                   | 2.1839 E-02    |

**Table 2.**  
Model with strong gradient: Error in  $\|\cdot\|_{L^2}$  for  $h = (\frac{1}{2})^4$ .

| $p$ | Enriched mixed ( $p$ ) | Mixed ( $p + 1$ ) | $H^1$ -conforming ( $p + 2$ ) | DG ( $p + 2$ ) |
|-----|------------------------|-------------------|-------------------------------|----------------|
| 2   | 2.6624 E + 00          | 2.6597 E + 00     | 2.9099 E + 00                 | 2.9832 E + 00  |
| 3   | 1.3588 E + 00          | 1.3568 E + 00     | 1.4473 E + 00                 | 1.4713 E + 00  |
| 4   | 8.8127 E-01            | 8.8046 E-01       | 9.754 E-01                    | 9.8129 E-01    |
| 5   | 4.70638 E-01           | 4.6996 E-01       | 5.1467 E-01                   | 5.2325 E-01    |
| 6   | 2.72517 E-01           | 2.7236 E-01       | 3.1024 E-01                   | 3.1815 E-01    |
| 7   | 1.6494 E-01            | 1.6479 E-01       | 1.9155 E-01                   | 1.9596 E-01    |

**Table 3.**  
Model with strong gradient: Error in  $\|\cdot\|_{H^1}$  for  $h = (\frac{1}{2})^4$ .

| $p$ | Mixed ( $p + 1$ )<br>clock ticks | Enriched mixed ( $p$ )<br>percent over mixed | $H^1$ -conforming ( $p + 2$ )<br>percent over mixed | DG ( $p + 2$ )<br>percent over mixed |
|-----|----------------------------------|----------------------------------------------|-----------------------------------------------------|--------------------------------------|
| 2   | 36.717.128                       | 69%                                          | 17%                                                 | 189%                                 |
| 3   | 61.787.577                       | 75%                                          | 19%                                                 | 288%                                 |
| 4   | 103.659.645                      | 83%                                          | 23%                                                 | 446%                                 |
| 5   | 184.153.128                      | 86%                                          | 24%                                                 | 529%                                 |
| 6   | 317.502.111                      | 88%                                          | 27%                                                 | 557%                                 |
| 7   | 536.418.532                      | 92%                                          | 29%                                                 | 589%                                 |

**Table 4.**

Comparing the computational cost of time processing for different formulations with that of the mixed method.

Given these considerations, we excluded the discontinuous Galerkin formulation and validated the remaining three formulations for both a 3D extension and a 2D scenario exhibiting highly oscillatory behavior.

### 3.3 Three-dimensional Poisson model exhibiting a strong gradient in the exact solution

Let us examine the 3D scenario to verify the findings derived from the 2D instance with a pronounced gradient, as discussed in Section 3.4. In this 3D scenario, we have excluded the discontinuous Galerkin formulation due to the high number of DoFs and the inability to implement static condensation.

Examine the Poisson model given by Eqs. (1) and (2) with the subsequent exact solution:

$$u_{exact}(x, y, z) = \left(\frac{2}{5}\right) \left[\frac{\pi}{2} + \arctan[10 - 200(x^2 + y^2 + z^2)]\right], \quad (15)$$

The source function  $f$  corresponds to  $f = -\Delta u_{exact}$  in the domain  $\Omega = [-1, 1]^3$ , according to Dirichlet boundary conditions. High gradients occur near the sphere centered at the origin with a radius of  $\frac{1}{4}$ . The exact solutions for pressure and flow are illustrated in **Figures 6** and **7**, respectively.

**Table 5** presents the errors in the  $||\cdot||_{L^2}$  norm derived from the three formulations for the strong gradient model. Examine the proximity of the mistakes when contrasting the traditional FEM approach of order ( $p + 2$ ), the mixed method of order ( $p + 1$ ), and the enriched mixed method of order  $p$

**Table 6** presents the errors in the  $||\cdot||_{H^1}$  norm for the 3D model exhibiting strong gradient. The approximation of the mixed approaches about the  $||\cdot||_{H^1}$  norm is highly accurate. The mixed methods approach is optimal. The classical  $H^1$ -conforming formulation is no longer as approximate as the mixed ones.

**Table 7** comparison illustrates the computational cost of the time processing (clock ticks) of the enriched mixed method and  $H^1$ -conforming in relation to the time utilized by the mixed method. Our observations indicate that the enriched mixed method is significantly less expensive than the mixed finite element method, yet the resulting errors were remarkably comparable, as demonstrated in **Table 6**. We will confirm that the same behavior is observed in the case of a highly oscillatory solution in the subsequent section.

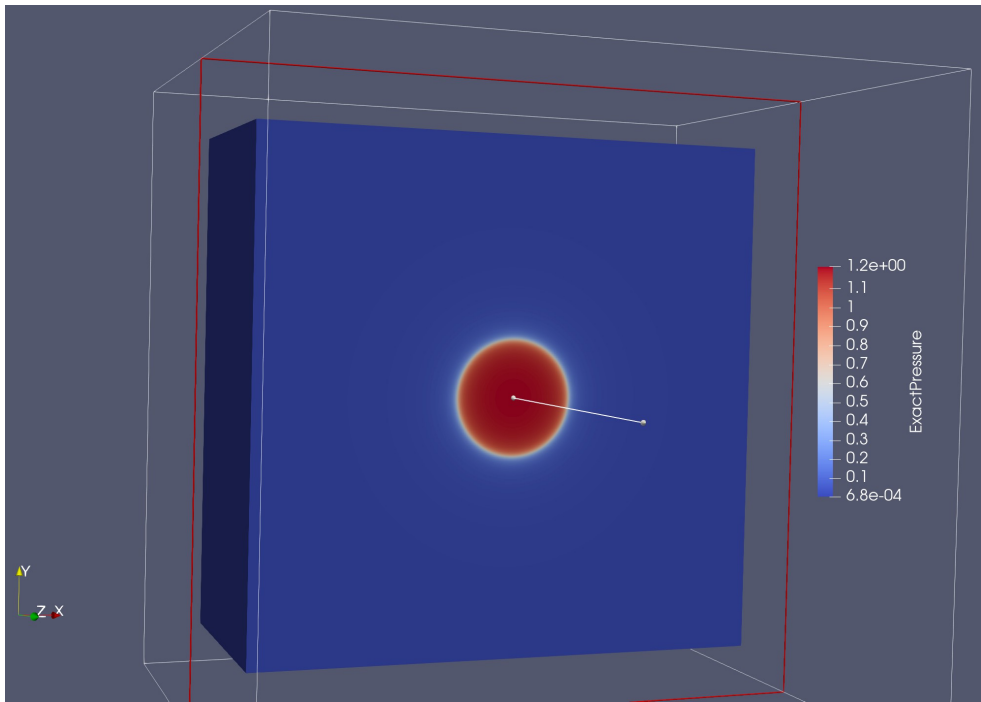


**Figure 6.**  
*The exact solution given by Eq. (15). The solution was bisected to illustrate the strong gradient.*

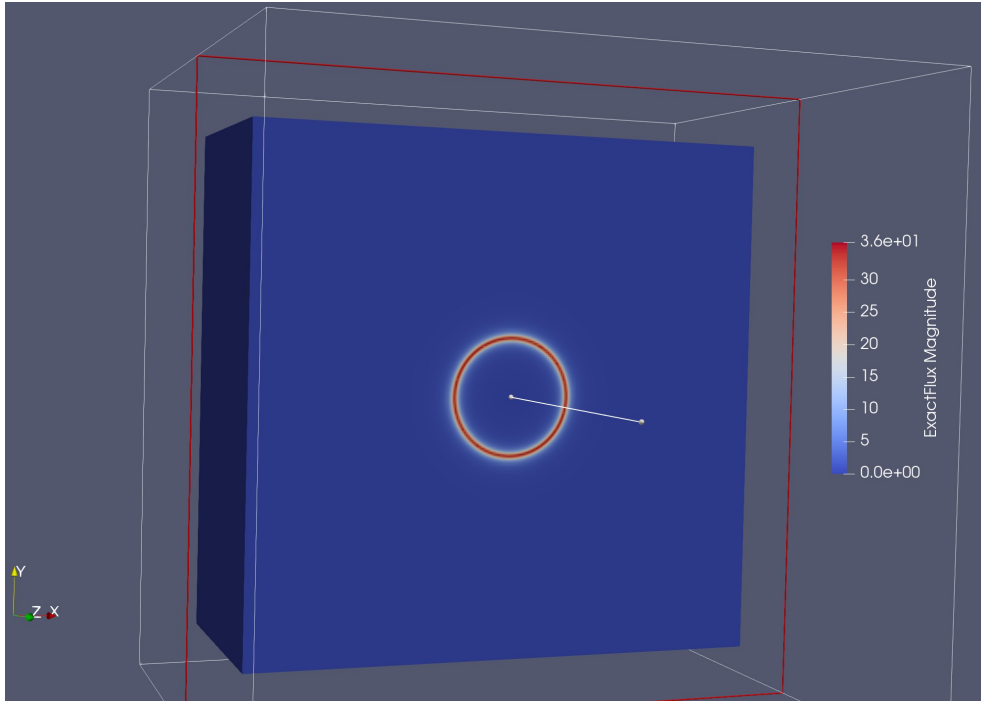


**Figure 7.**  
*The exact flux of the solution given by Eq. (15). The solution was bisected to exhibit the flux near the sphere.*

| $p$ | Enriched mixed ( $p$ ) | Mixed ( $p + 1$ ) | $H^1$ -conforming ( $p + 2$ ) |
|-----|------------------------|-------------------|-------------------------------|
| 1   | 1.9908 E + 00          | 1.9900 E + 00     | 1.9883 E + 00                 |
| 2   | 1.9891 E + 00          | 1.9890 E + 00     | 1.9899 E + 00                 |
| 3   | 4.6802 E-01            | 4.6794 E-01       | 4.6943 E-01                   |
| 4   | 1.8186 E-01            | 1.8179 E-01       | 1.8268 E-01                   |
| 5   | 5.1361 E-01            | 5.1361 E-01       | 5.1359 E-01                   |

**Table 5.**  
3D model with strong gradient: Error in  $||\cdot||_{L^2}$  for  $h = (\frac{1}{2})^3$ .

| $p$ | Enriched mixed ( $p$ ) | Mixed ( $p + 1$ ) | $H^1$ -conforming ( $p + 2$ ) |
|-----|------------------------|-------------------|-------------------------------|
| 1   | 8.38102                | 8.33697           | 8.60069                       |
| 2   | 8.53888                | 8.53727           | 9.18953                       |
| 3   | 3.76435                | 3.75397           | 4.38396                       |
| 4   | 3.66037                | 3.65564           | 4.08231                       |
| 5   | 3.66837                | 3.66809           | 3.78815                       |

**Table 6.**  
3D model with strong gradient: Error in  $||\cdot||_{H^1}$  for  $h = (\frac{1}{2})^3$ .

| $p$ | Mixed ( $p + 1$ ) clock ticks | Enriched mixed ( $p$ ) percent over mixed | $H^1$ -conforming ( $p + 2$ ) percent over mixed |
|-----|-------------------------------|-------------------------------------------|--------------------------------------------------|
| 1   | 543.897.277                   | 36%                                       | 13%                                              |
| 2   | 2.626.818.995                 | 49%                                       | 5%                                               |
| 3   | 13.130.677.152                | 65%                                       | 6%                                               |
| 4   | 20.360.581.174                | 77%                                       | 7%                                               |
| 5   | 184.612.803.119               | 80%                                       | 10%                                              |

**Table 7.**  
Comparing the computational cost of time processing for different formulations with that of the mixed method.

### 3.4 2D Poisson model exhibiting a highly oscillatory exact solution

Examine the Poisson model given by Eqs. (1) and (2) with the subsequent exact solution:

$$u_{exact}(x,y) = \left(\frac{2}{5}\right)[\sin(9\pi x)][1 + \cos(9\pi y)]\left[\frac{1}{2}\pi + \arctan[10 - 200(x^2 + y^2)]\right], \quad (16)$$

The function  $f$  corresponds to  $f = -\Delta u_{exact}$  in the domain  $\Omega = [-1, 1]^2$ , according to Dirichlet boundary conditions. Highly oscillatory behavior is seen within the circle

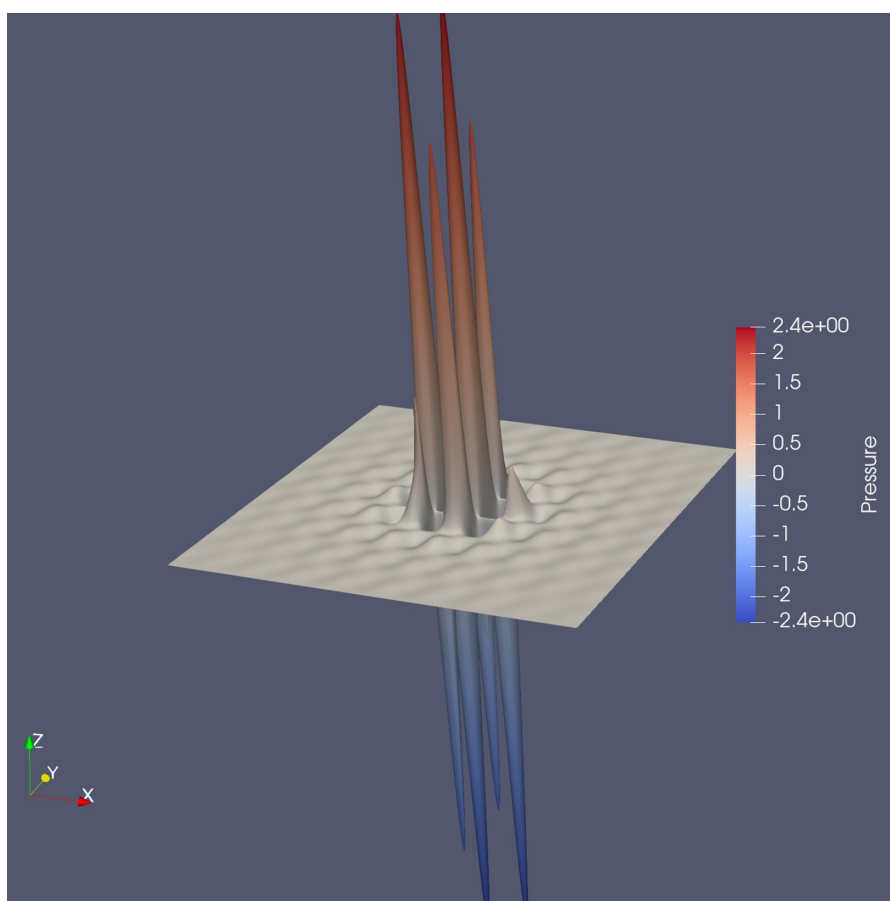


**Figure 8.**  
 The exact solution given by Eq. (16) exhibits highly oscillatory behavior.

centered at the origin, with a radius of  $\frac{1}{4}$ . The exact solutions for pressure and flow are illustrated in **Figures 8** and **9**, respectively.

**Table 8** shows the very close errors in the  $\|\cdot\|_{L^2}$  norm obtained for the three formulations for the highly oscillatory model. **Figure 10** shows a plot of the errors.

**Table 9** presents the errors in the  $\|\cdot\|_{H^1}$  norm for the highly oscillatory model. Similar to other instances, the mixed techniques regarding the  $\|\cdot\|_{H^1}$  norm exhibit a high degree of proximity. The classical  $H^1$ -conforming is no longer as approximate as the mixed ones. **Figure 11** shows a plot of the errors.

## 4. Conclusions

The enhanced mixed finite element methodology was confirmed to meet the theoretical convergence criteria. The enrichment mixed method of order  $p$  has demonstrated an approximation akin to that of the mixed finite element method of order  $(p + 1)$ . Nevertheless, its computational expense is reduced by as much as 67%, since it accounts for fewer DoFs owing to a diminished quantity of vector

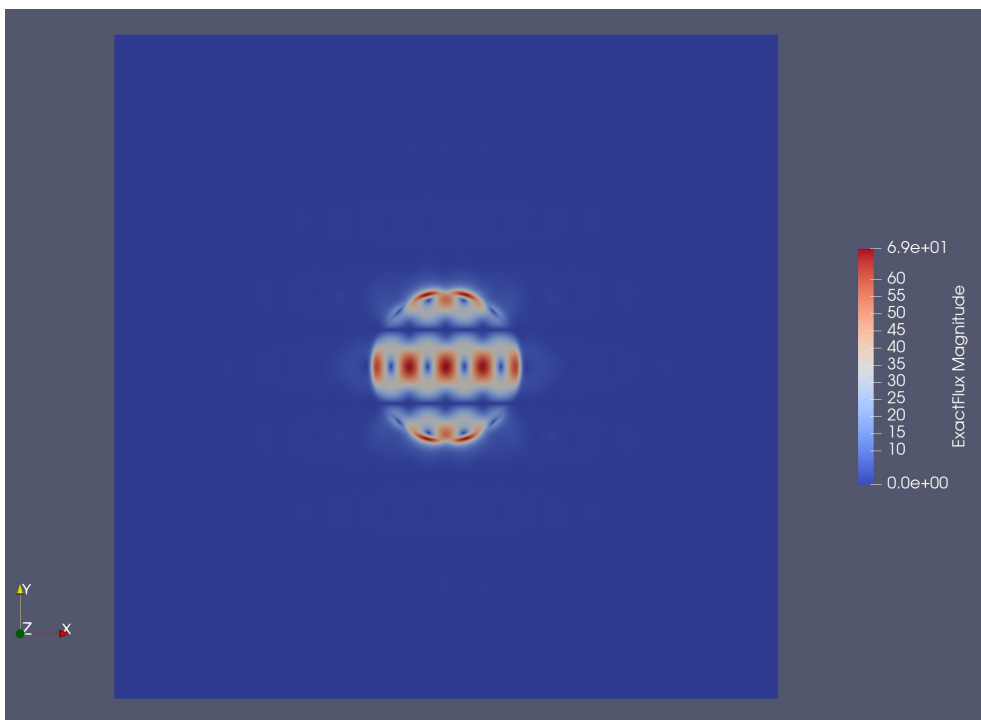


**Figure 9.**  
The exact flux of the solution given by Eq. (16).

| $p$ | Enriched mixed ( $p$ ) | Mixed ( $p + 1$ ) | $H^1$ -conforming ( $p + 2$ ) |
|-----|------------------------|-------------------|-------------------------------|
| 2   | 3.6501 E-02            | 3.6503 E-02       | 3.6458 E-02                   |
| 3   | 2.8593 E-02            | 2.8595 E-02       | 2.8421 E-02                   |
| 4   | 6.7319 E-03            | 6.7340 E-03       | 6.6467 E-03                   |
| 5   | 5.3172 E-03            | 5.3180 E-03       | 5.2985 E-03                   |
| 6   | 4.9920 E-03            | 4.9920 E-03       | 4.8741 E-03                   |
| 7   | 3.4295 E-03            | 3.4290 E-03       | 3.4059 E-03                   |

**Table 8.**  
Highly oscillatory model: Error in  $\|\cdot\|_{L^2}$  for  $h = (\frac{1}{2})^4$ .

functions linked to the edges of the master element. The enriched version signifies an optimized iteration of mixed finite element methods. The traditional finite element method demonstrates markedly lower computational expenses, even for orders  $(p + 2)$  or above, while delivering enhanced approximation in the  $\|\cdot\|_{L^2}$  norm of the primary variable. Nonetheless, it is recognized that the representation of indirectly derived flows in the classical finite element method is of substandard quality.

Given that the errors in the  $\|\cdot\|_{L^2}$  norm (which are quite similar) and the  $\|\cdot\|_{H^1}$  norm are comparable, we emphasize that the mixed enriched formulation of degree  $p$  is equivalent to the mixed formulation of degree  $(p + 1)$ , and it is also analogous to the traditional FEM formulation of degree  $(p + 2)$ .

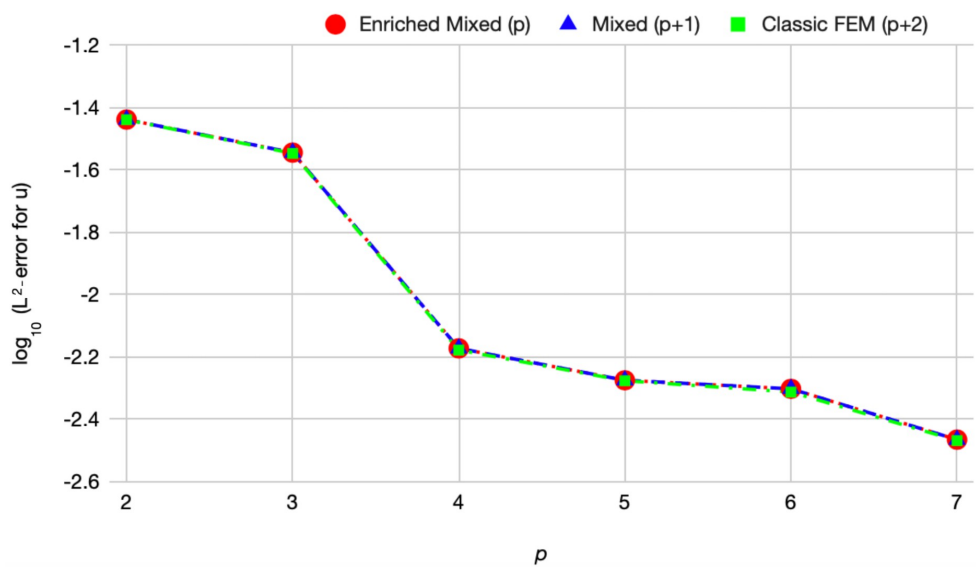


**Figure 10.**  
Pressure variable: Error in  $\|\cdot\|_{L^2}$  for the strongly oscillatory model with  $h = (\frac{1}{2})^4$ .

| $p$ | Enriched mixed ( $p$ ) | Mixed ( $p + 1$ ) | $H^1$ -conforming ( $p + 2$ ) |
|-----|------------------------|-------------------|-------------------------------|
| 2   | 1.4551 E + 00          | 1.4350 E + 00     | 1.6868 E + 00                 |
| 3   | 9.4273 E-01            | 9.3923 E-01       | 1.044 E + 00                  |
| 4   | 6.0902 E-01            | 6.0810 E-01       | 7.1658 E-01                   |
| 5   | 4.098 E-01             | 4.0876 E-01       | 4.6148 E-01                   |
| 6   | 2.7115 E-01            | 2.7098 E-01       | 3.2247 E-01                   |
| 7   | 1.8533 E-01            | 1.8465 E-01       | 2.1034 E-01                   |

**Table 9.**  
Highly oscillatory model: Error in  $\|\cdot\|_{H^1}$  for  $h = (\frac{1}{2})^4$ .

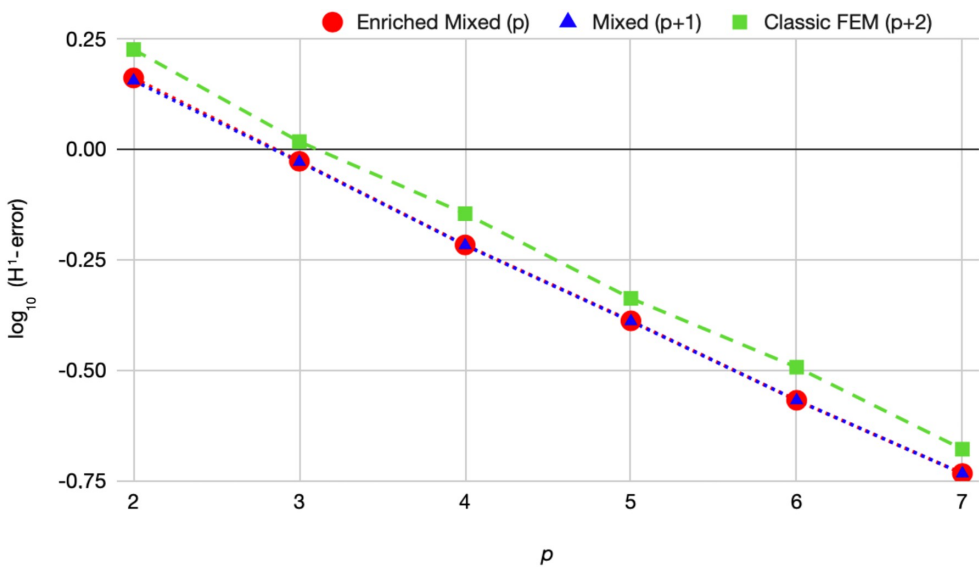


**Figure 11.**  
Error in norm  $H^1$  for the strongly oscillatory model with  $h = (\frac{1}{2})^4$ .

### Author details

Jorge L. Díaz Calle<sup>1\*</sup>, Roger Mestas Chávez<sup>2\*</sup> and Aldo Sarmiento Guevara<sup>3\*</sup>

1 Department of Basic Sciences - FZEA, University of São Paulo, Campus Fernando Costa da USP, Pirassununga, Brazil

2 Universidad Nacional de San Agustín de Arequipa y Universidad Católica de Santa María, Arequipa, Perú

3 Universidad Nacional de San Agustín de Arequipa, Arequipa, Perú

\*Address all correspondence to: jorge.calle@usp.br; rmestas@ucsm.edu.pe; asarmientog@unsa.edu.pe

### IntechOpen

© 2025 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Sarmiento A, Aynayanque J, Condori R, Diaz J. Study of the enriched mixed finite element method using comparisons of computational cost and errors with formulations in continuous and discontinuous functions and mixed scheme on quadrilateral finite elements. In: Proceedings of the XLI Ibero-Latin-American Congress on Computational Methods in Engineering; *ABMEC*; 2020
- [2] Hanco R, Aynayanque J, Condori R, Velasquez E, Mestas R, Mamani F, Diaz J. Comparative analysis of the performance of a enriched mixed finite element method with static condensation for poisson problems. In: Proceedings of the XLI Ibero-Latin-American Congress on Computational Methods in Engineering, *ABMEC*; 2020.
- [3] Brezzi F, Fortin M. Mixed and hybrid finite element methods. In: Springer Series in Computational Mathematics. Vol. 15. New York: Springer-Verlag; 1991
- [4] Raviart AP, Thomas JM. A mixed finite element method for 2-nd order elliptic problems. In: Galligani I, Magenes E, editors. *Mathematical Aspects of Finite Element Methods. Lectures Notes in Mathematics*. 1977. p. 292–315
- [5] Nedelec JC. A new family of mixed finite elements in  $r_3$ . *Numerische Mathematik*. 1986;**50**:57–81
- [6] de Siqueira D, Devloo PRB, Gomes SM. A new procedure for the construction of hierarchical high order  $h_{div}$  and  $h_{curl}$  finite element spaces. *Journal of Computational and Applied Mathematics*. 2016;**240**:204–214
- [7] Devloo PRB, Gomes SM, Quinelato TO, Tian S. Enriched two dimensional mixed finite element models for linear elasticity with weak stress symmetry. *Computers and Mathematics with Applications*. 2020;**79**:2678–2700
- [8] Calle JL, Pastor J, Bustincio R, Ancori R. Study of the enriched mixed finite element method comparing errors and computational cost with classical fem and mixed scheme on quadrilateral meshes. *Results in Applied Mathematics*. 2021;**10**:100150
- [9] Baumann CE. An Hp-adaptive Discontinuous Finite Element Method for Computational Fluid Dynamics. PhD thesis, University of Texas at Austin, Texas, USA, 1997
- [10] Farias AM, Devloo PRB, Gomes SM, de Siqueira D, Castro DA. Two dimensional mixed finite element approximations for elliptic problems with enhanced accuracy for the potential and flux divergence. *Computers and Mathematics with Applications*. 2017;**74**:3283–3295
- [11] Forti TLD, Farias AM, Devloo PRB, Gomes SM. A comparative numerical study of different finite element formulations for 2d model elliptic problems: Continuous and discontinuous galerkin, mixed and hybrid methods. *Finite Elements in Analysis and Design*. 2016;**115**:9–20
- [12] Devloo PRB. PZ: An object oriented environment for scientific programming. *Computer Methods in Applied Mechanics and Engineering*. 1997;**150** (1–4):133–153

[13] Bravo CM, Pavanello R, Devloo PR, JI C. Definition of a p-interpolating space of hierarchical bases of finite elements on the pyramid. *Linear Algebra and Its Applications*. 2014;**460**:174–204

[14] Devloo PRB, Bravo CMAA, Rylo EC. Systematic and generic construction of shape functions for p-adaptive meshes of multidimensional finite elements. *Computer Methods in Applied Mechanics and Engineering*. 2009;**198**:1716–1725

[15] Castro DA, Devloo PRB, Farias AM, Gomes SM, de Siqueira D, Duran O. Three dimensional hierarchical mixed finite element approximations with enhanced primal variable accuracy. *Computer Methods in Applied Mechanics and Engineering*. 2016;**306**:479–502

[16] Devloo P, Gomes S, Quinelato T, Tian S. Enriched two dimensional mixed finite element models for linear elasticity with weak stress symmetry. *Computers and Mathematics with Applications*. 2020;**79**:2678–2700

---

## Section 3

# Applications in Engineering and Science

---



## Chapter 4

# FEM Modeling Techniques for Industrial Applications

*Francisco Badea-Romero and Jesús A. Perez-Fernandez*

### Abstract

The finite element method (FEM) is a core tool in modern engineering, enabling structural evaluation before prototyping. In industries, its application is shaped by cost, time, and reliability, leading to streamlined modeling strategies. This chapter reviews practical FEM techniques for mechanical and manufacturing sectors, focusing on geometry preparation, element selection, connection modeling, meshing strategies, and material definition. Industrial practice favors beams and shells for efficiency, with solids, springs, dampers, and mass elements applied selectively. Connections are often simplified using rigid body elements (RBEs), multi-point constraints (MPCs), or contact elements, supported by post-processing. Meshes emphasize quadrilateral shells, hybrid solid meshes, and first-order elements. Materials are typically linear elastic, with nonlinear models reserved for critical cases. The chapter offers best-practice guidelines for achieving accurate, cost-effective FEM simulations under industrial constraints.

**Keywords:** finite element method (FEM), industrial simulation practices, element selection, meshing strategies, material modeling, connection modeling, structural analysis

### 1. Introduction

The Finite Element Method (FEM) has become one of the cornerstones of modern engineering, providing an indispensable tool for analyzing, optimizing, and validating designs before physical prototyping. Over the past few decades, constant advancements in computational power, software capabilities, and numerical techniques have significantly expanded the scope and accuracy of FEM simulations. A recent historical overview by Liu et al. [1] traces how this method has transformed from early formulations to its current role as a fundamental tool across many industrial applications.

In industrial environments, however, the theoretical potential of FEM is only part of the equation. Engineers often operate under strict constraints of time, cost, and manufacturability. In this context, simulation-driven design, where FEM tools guide design decisions and virtual prototypes replace physical iterations, has emerged as a powerful and practical strategy to accelerate project timelines in industrial

environments [2]. Consequently, a set of unspoken rules and practical approaches exists – methods that may not always be explicitly documented in textbooks but which consistently prove to be effective, reliable, and cost-efficient in day-to-day operations.

This chapter presents a detailed overview of these commonly applied industrial practices in the mechanical sector. The discussion will focus on strategies for model preparation, meshing, material definition, load application, and result validation, with emphasis on balancing computational efficiency and engineering accuracy. Selected examples from mechanical and manufacturing industries will illustrate how these practices can be implemented to achieve optimal results with minimal resource expenditure.

## **2. Geometry preparation and model simplification**

### **2.1 Importance of pre-processing in industry**

In an industrial context, the efficiency and reliability of an FEM simulation often depend less on raw computational power and more on the quality of the pre-processing stage. Geometry imported directly from CAD is almost never ready for immediate simulation. While the CAD model serves as the authoritative representation for manufacturing, it typically contains fine details, small features, manufacturing tolerances, and decorative elements that may be critical for production but have negligible influence on the FEM model.

These extra details, such as embossed logos, small-radius fillets, small holes, or surface textures, are completely unnecessary for analysis purposes. They can significantly increase meshing complexity without adding value to the simulation and, in some cases, may even make meshing or analysis impossible. For this reason, an essential first step in industrial FEM is geometry simplification and defeaturing, ensuring that only features influencing the behavior under the studied load cases are retained.

It is also important to note that element-type selection is almost never performed by CAD designers. Their focus is on manufacturability and dimensional accuracy, not on numerical efficiency or stress resolution. The responsibility for deciding whether to use 1D beams, 2D shells, or 3D solids lies with the simulation engineer, who understands both the mechanical problem and the trade-offs between computational cost and accuracy. This separation of roles is a common industrial practice, ensuring that FEM models are tailored to engineering requirements rather than CAD esthetics.

### **2.2 Common industrial practices**

In the industrial application of FEM, certain pre-processing techniques have proven to be both widely used and highly effective in improving simulation efficiency. Among these, defeaturing and model adaptation are the most relevant, while symmetry exploitation is less frequently applied. While these practices are conceptually straightforward, their implementation requires experience, engineering judgment, and meticulous attention to detail.

### 2.2.1 Defeaturing

Defeaturing, the removal of unnecessary geometric details from the CAD model, is by far the most common industrial pre-processing practice. Its purpose is to simplify the geometry without compromising the accuracy of the simulation. The specific defeaturing steps vary depending on the CAD software used, as each offers different tools and levels of automation. However, the most frequently applied modifications include:

- Removal of small fillets and rounds that do not significantly influence the stress or thermal distribution.
- Patching small holes that would create unnecessary mesh refinement.
- Correction of gaps between components, such as welding gaps, ensures continuous load transfer in the simulation.

An example of geometry defeaturing is shown in **Figure 1**, where rounds, fillets, gaps, and esthetic features with no structural relevance are removed. It is important to note that one fillet radius has been retained, illustrating the preservation of features with potential influence on the calculation results. In this case, the radius may affect stress concentration and, therefore, the peak stress values.

These operations often require manual or semi-manual intervention, with the engineer inspecting the model piece by piece to ensure that important features are preserved.

### 2.2.2 Model adaptation

Depending on the nature of the structure, the FEM model may require significant adaptation beyond defeaturing. For example:

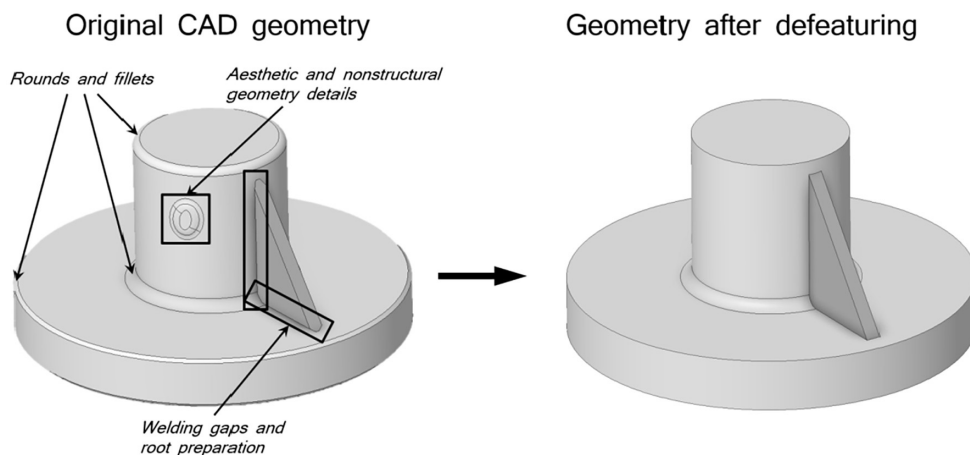


**Figure 1.**  
*Geometry defeaturing example.*

- In tubular structures, all 3D beam solids from the CAD model are often discarded and replaced with keypoints and lines that will be used as guides for later 1D beam elements.
- When shell-type elements are chosen, mid-surface extraction is necessary. This operation, depending on the characteristics of the available CAD or FEM software, can be done semi-automatically or manually and can take significant amounts of time. Additionally, modifications such as artificially extending components so that their mid-surfaces maintain the intended contact are required.

**Figure 2** illustrates a simple example of model adaptation involving beam-type and shell-type geometries. As a rule of thumb, in a three-dimensional Cartesian reference system, beam-type geometries are characterized by one dimension being significantly larger than the other two, while shell-type geometries are characterized by two dimensions being significantly larger than the third. In the adapted model, the junction between different components is no longer represented as a contact surface but rather as a single point. Consequently, the local stiffness of the joint is inherently modified during this process, as demonstrated in [3].

A decision must be made whether contacting shells should be merged into a single shell with combined thickness or kept as two separate shells with their actual thicknesses, connected via contact elements. This choice depends on accuracy requirements, contact stiffness considerations, and post-processing needs.

### 3. Element type selection and modeling characteristics

#### 3.1 Importance of element selection in industry

In FEM, the choice of element type directly influences both the accuracy of results and the computational cost of the simulation. While textbooks present element selection as a purely technical decision based on geometry, boundary

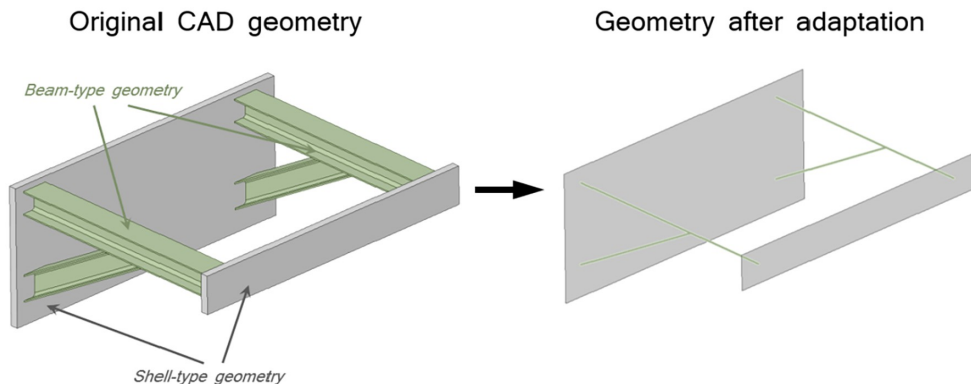


**Figure 2.**  
*Geometry adaptation example.*

conditions, and physics, industrial practice adds further constraints such as project deadlines, licensing limitations, available computational resources, and the need for repeatable, efficient workflows.

In industrial environments, the main determining factor is cost, which is inherently tied to time, engineering effort, and project deadlines. This reality often forces trade-offs between theoretical precision and practical deliverables, with engineers selecting the simplest element configuration that meets the accuracy requirements while minimizing computational expense.

### **3.2 Main element selection in the industry**

#### *3.2.1 Beam elements*

Beam-type elements are among the most simplified finite element formulations, defined by two nodes and governed by the mathematical models of Euler–Bernoulli or Timoshenko beam theory. While these elements significantly simplify the geometric characteristics of beams, particularly at joints and connections, they are not capable of providing detailed local stress or strain distributions, especially in complex joint regions or at welds.

Despite these limitations, beam elements remain a mainstay in industrial simulations whenever large tubular or frame-type structures need to be analyzed efficiently. Their low computational cost and ability to capture overall stiffness and deflection behavior make them ideal for applications such as:

- Upper structures of buses and coaches.
- Solar panel support structures.
- Long-span bridges.
- Tower cranes and gantry cranes.
- Large steel frames and truss systems.

**Figure 3** shows some characteristic examples of a recent bibliography that illustrates typical applications of beam-type elements to finite element structural calculations.

It is worth noting that, despite significant software advancements that allow tubular structures to be modeled with shell-type elements for higher geometric fidelity, the complexity of the modeling process and the time required for geometry preparation, mid-surface extraction, and connection detailing make shell modeling impractical for most industrial timelines. As a result, beam elements remain the go-to choice for large tubular assemblies, striking the best balance between accuracy, speed, and resource efficiency in industrial environments.

#### *3.2.2 Shell elements*

Shell elements are by far the most widely used element type in industrial FEM simulations. Despite theoretical guidelines suggesting other approaches for certain

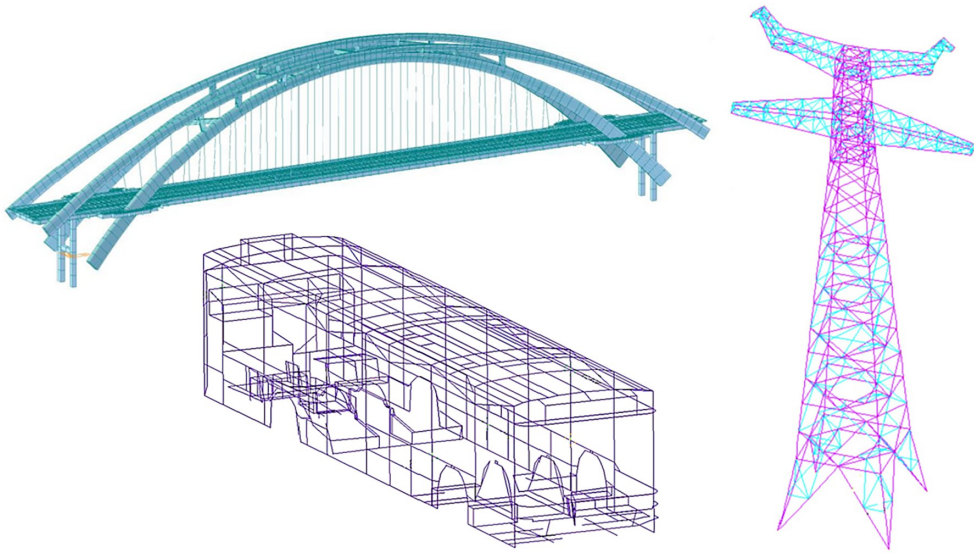


**Figure 3.** Examples of beam-type elements in finite element models. Top-left: long span bridge [4]. Bottom-left: passenger transport urban bus [5]. Right: large transmission tower[6].

cases, shells dominate in sectors such as aerospace, railway, automotive, and general industrial machinery. **Figure 4** shows some characteristic examples of shell element-type utilization in actual industrial applications.

Their popularity stems from their ability to represent complex, thin-walled structures with a good balance between accuracy and computational efficiency, making them suitable for the vast majority of structural analyses performed in the industry.

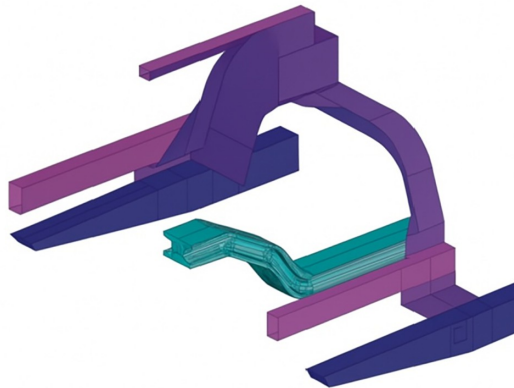


**Figure 4.** Examples of shell-type elements in the finite element model for a bus chassis, left: reference structure; right: FEM shell model.

From a mathematical standpoint, shell elements are derived from classical plate and shell theories, most notably Kirchhoff–Love (for thin shells, neglecting transverse shear deformation) and Mindlin–Reissner (for moderately thick shells, including shear effects). These formulations allow shell elements to capture both in-plane membrane and out-of-plane bending behaviors, making them versatile for a wide range of loading scenarios.

While they do require significant pre-processing and experience to use effectively, including mid-surface extraction, proper thickness assignment, and careful definition of connections or overlaps, shell elements provide a sufficiently accurate geometric representation of most industrial structures at a manageable computational cost. This combination of fidelity and efficiency explains why they are chosen even when other types of elements might theoretically offer higher accuracy.

Like all element types, shells still simplify the real structure and have certain limitations:

- They assume a constant thickness over each element.
- Through-thickness stress variations are not captured in detail.
- Local 3D effects at joints or discontinuities require complementary modeling (e.g., solid submodels).

Nevertheless, shells offer the best trade-off between required accuracy, modeling time, and computational resources in industrial practice, making them the default choice for many engineering teams.

### *3.2.3 Solid elements*

Solid elements, also known as 3D continuum elements, are among the easiest to apply in FEM from a modeling perspective, as they require no mid-surface extraction or special geometric simplifications. However, despite their ease of use, they are one of the least utilized element types in industrial practice. Their use is generally confined to complex block-type geometries that cannot be effectively simplified into beam or shell representations. Typical examples include machined or forged components, as well as detailed simulations of biomedical or organic structures, such as the ones illustrated in **Figure 5**.

This limited usage is primarily due to their high computational cost; accurate simulations with solid elements require fine meshes to capture stress gradients, leading to large numbers of elements and correspondingly longer solution times. Although any structure that can be modeled with beam or shell elements could, in principle, be modeled using solids, in practice, this is almost never done because of the large resource requirements.

From a mathematical perspective, solid elements are based on 3D elasticity theory, supporting full-displacement fields and capable of representing complete stress and strain tensors throughout the volume of the component.

A critical factor when working with solid elements in industry is mesh quality. Poorly shaped elements can introduce numerical inaccuracies or cause convergence problems. The most important quality parameters are:

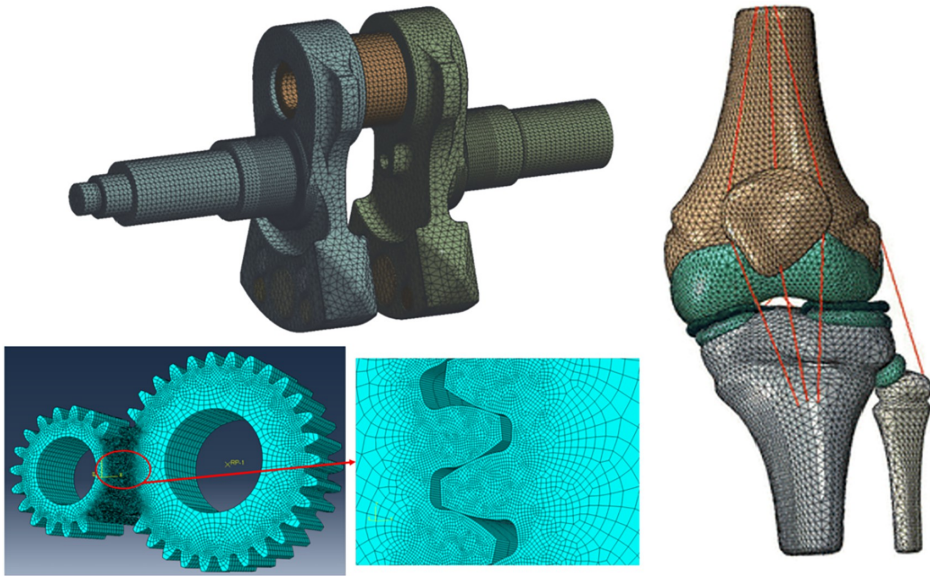


**Figure 5.**  
*Examples of solid elements in finite element models. Top-left: engine crankshaft [8]. Right: knee joint [9]. Bottom-left: gear transmission [10].*

- Skewness: Excessive skew in element geometry can distort the displacement field and reduce accuracy.
- Aspect ratio: Large differences in edge lengths should be avoided, except in low-stress regions, to prevent stiffness errors.

In industrial workflows, solid elements are typically reserved for:

- Bulky components that cannot be represented by shell elements, such as castings, thick brackets, or solid machine parts.
- Localized submodels, where through-thickness stress variations are critical.
- Research and development tasks, where maximum geometric fidelity is prioritized over computational cost.

While solids are conceptually straightforward and versatile, their computational demands and sensitivity to mesh quality make them the least-used element type in most industrial structural simulations, where beams and shells dominate due to their efficiency and practicality.

### *3.2.4 Spring elements*

Spring elements are specialized finite element formulations used to simulate elastic behavior between two points or components. They can be applied in a translational or rotational form, depending on whether they represent linear or angular stiffness.

Although a wide variety of mathematical models are available, including nonlinear elastic springs, viscoelastic springs, and frequency-dependent spring–damper systems, in industrial practice, the vast majority of applications use linear elastic spring elements. This is primarily because linear springs are computationally inexpensive, require no iterative solution procedures, and can be easily implemented without complex parameter definitions.

Nonlinear elastic springs, though theoretically more accurate for certain materials or connections, are rarely used in industrial workflows. Their use requires iterative simulations, which, in turn, increase computation time and resource consumption, making them impractical for most production schedules.

From a modeling standpoint, spring elements are most often implemented as single degree of freedom (1DOF) connectors, acting in either tension/compression or rotation. This 1DOF nature adds to their simplicity and versatility, allowing them to be inserted into large models without introducing unnecessary complexity.

Common industrial applications of spring elements include:

- Modeling actual mechanical springs.
- Representing flexible mounts, supports, or isolators.
- Simulating the stiffness of gaskets or compliant layers without detailed geometry.
- In some cases, modeling bolted or riveted joints as elastic connectors are used to approximate joint compliance.

Spring elements provide a fast and efficient method to include elastic behavior in an FEM model, especially when the goal is to capture global structural response without investing time and resources in detailed component modeling.

### *3.2.5 Damper elements*

In most FEM software packages, damper elements share the same underlying formulation as spring elements, with the primary difference being the replacement of stiffness with a damping coefficient. From a modeling standpoint, defining a damper element is often as simple as setting the stiffness to zero and assigning a nonzero damping coefficient.

A notable feature is that these elements can also act as:

- Pure springs if the damping coefficient is set to zero.
- Pure dampers if the stiffness is set to zero.
- Spring–damper combinations if both stiffness and damping coefficients are nonzero, allowing them to represent systems that store and dissipate energy simultaneously.

Damper elements only contribute to the system response in dynamic analyses, such as transient simulations or frequency-response studies. In purely static

simulations, which represent the majority of industrial mechanical analyses, damper elements have no effect because velocities are zero.

In industrial contexts, damper elements are used far less frequently than springs because static analysis dominates most mechanical engineering workflows.

When they are applied, damper elements typically appear in:

- Vibration isolation systems in machinery.
- Shock absorber and suspension system modeling in the automotive sector.
- Impact simulations where energy dissipation affects structural response.
- Rotating machinery simulations to study dynamic stability and reduce resonances.

In summary, damper elements are easy to implement, versatile in their configurations, and can model pure damping, pure elasticity, or combined spring–damper behavior. However, their use in industries is limited to problems where dynamic effects and energy dissipation are critical considerations.

### *3.2.6 Mass elements connected with RBE*

In many industrial simulations, non-structural components, such as instrumentation, auxiliary equipment, or payloads, do not require detailed modeling of their geometry or stiffness but still contribute significantly to the overall mass distribution of the structure. In these cases, the most common practice is to represent such components using a lumped mass element with the corresponding total mass and connect it to the structural model using simple rigid body elements (RBEs).

The lumped mass element is typically placed at the center of mass of the component, and the RBE connection is defined between this mass node and a location on the structural model that best represents the real attachment point. This approach greatly reduces modeling complexity and computational cost while preserving the essential inertial effects in dynamic analyses.

Several RBE types can be used for the connection, each with its own advantages and limitations:

- Rigid Element (generic): Transfers all degrees of freedom without flexibility.
- RBE1: Distributes motion from independent nodes to dependent nodes while maintaining relative displacement.
- RBE2: Enforces rigid motion between connected nodes, distributing forces evenly.
- RBE3: Transfers loads and displacements without adding stiffness to the system.

The choice of RBE type is usually made by the simulation engineer based on experience, the specific objectives of the analysis, and the desired balance between stiffness representation and load transfer fidelity.

While this method is efficient, it can introduce local deviations in results due to the artificial stiffness or load distribution introduced by the RBE connection.

The most common industrial practice is to place the lumped mass element at the component's mass centre and connect it via the most realistic RBE configuration, avoiding unnecessary complexity while maintaining accurate inertial representation for the intended simulation purpose.

### *3.2.7 RBE elements for the simulation of bolted and riveted connections*

In industrial FEM practice, bolted and riveted connections are often simplified using RBEs to connect two or more structural components at the joint location. This approach allows engineers to represent the load transfer mechanism of the connection without modeling the full bolt or rivet geometry, threads, contact surfaces, or preloads, which would require far more complex and computationally expensive simulations.

The most common procedure involves creating an RBE element (e.g., RBE2 or RBE3) instead of the real bolt or rivet. The choice of RBE type depends on the desired balance between stiffness fidelity and load transfer representation:

- RBE2: Enforces rigid motion between connected nodes and is typically used while simulating a very stiff connection.
- RBE3: Distributes loads without adding artificial stiffness; often preferred when the flexibility of the joint area needs to be maintained.

In terms of engineering evaluation, the FEM simulation typically serves to determine the global force or stress resultants acting at each connection. These values are then exported and used in post-processing calculations, either through specialized software or manual engineering formulas, to verify the capacity of the bolts or rivets according to relevant standards (e.g., EN, ISO, ASME, or company-specific design rules). This hybrid approach, combining simplified FEM representation with separate direct calculations, allows for accurate and certifiable connection verification without overburdening the simulation model.

### *3.2.8 RBE elements for the simulation of welds*

In industrial FEM workflows, welded connections can be represented using RBEs to link the nodes of two components that are intended to be welded but do not share a common mesh line. This situation often arises when the mesh topology of each component differs, or when the CAD geometry does not have a continuous edge suitable for direct node merging.

The RBE-based approach involves creating rigid connections between the relevant nodes along the intended weld path, transferring displacements and loads between the two parts. While this procedure can be tedious if performed manually, most modern FEM software offer tools or macros to automate the process, making it possible to connect large numbers of weld seams quickly, consistently, and reliably.

Once the components are connected via RBEs, the evaluation of the welds is not performed directly from the FEM results alone. Instead, the engineer extracts forces and stress values at the connected nodes during post-processing. These values are

then used in separate weld design calculations, either manually or via specialized software, to verify the weld capacity according to applicable standards [7] (e.g., EN ISO 5817, AWS D1.1, or company-specific specifications).

The RBE method for weld modeling offers significant advantages in terms of modeling speed, numerical stability, and compatibility with varying mesh topologies. However, it must be understood that this is a simplified global representation: it does not capture local stress gradients within the weld throat or detailed metallurgical effects, which require more advanced submodeling techniques if needed.

## **4. Advances and alternatives in bolt, rivet, and weld simulation**

Mainstream FEM programs are in constant evolution, regularly introducing new interfaces, modules, and automated tools to assist engineers in simulating structural connections such as bolts, rivets, and welds. These developments aim to reduce modeling time, improve accuracy, and make simulation workflows more user-friendly. However, each approach has its own advantages and disadvantages, and the optimal method often depends on the type of analysis, available resources, and required result detail.

In the following paragraphs, the most common alternatives are presented.

### **4.1 Bolt and rivet simulation using beam elements**

Several FEM programs allow the creation of bolts or rivets as beam-type elements between two nodes, with a defined radius and material characteristics.

Pros:

- Simple to implement and computationally efficient.
- Can represent the axial and bending stiffness of the fastener.
- Allows easy parameterization for quick design changes.
- Works well for large assemblies with many fasteners, where local stress detail is not required.

Cons:

- Does not capture local contact pressure or shear distribution in the joint.
- Thread geometry, preload, and friction effects are not modeled directly.
- Still requires detailed post-processing for the correct calculation of the bolted connection in accordance with applicable standards.

### **4.2 Bolt, rivet, and weld simulation through MPC elements**

Multi-point constraint (MPC) elements connect multiple nodes through mathematical constraint equations, similar in purpose to RBEs but with greater

flexibility in defining the degrees of freedom to be coupled. MPCs are often used to simulate the stiffness and load transfer of fasteners or welds without adding mass or excessive rigidity.

Pros:

- More control over which DOFs are constrained, allowing partial flexibility in the connection.
- No artificial mass or inertia is introduced.
- Suitable for cases where the connection is stiff in some directions but compliant in others.

Cons:

- Slightly more complex to define compared to standard RBEs.
- Still does not represent the local deformation of the fastener or weld material.
- The constraint equations may affect solution performance in very large models if overused.

#### **4.3 Bolt, rivet, and weld simulation through contact elements with all DOF restrictions**

Another approach is to simulate the joint using contact elements between the connected surfaces, with all degrees of freedom (DOFs) restricted to mimic a welded or bolted connection.

Pros:

- Captures local contact pressure distribution and realistic load paths.
- Can be combined with preload or friction modeling for more realistic simulations.

Cons:

- More complex to set up compared to the other methodologies.
- Greater potential for convergence issues exists if contact parameters are not tuned correctly.

While beam, MPC, and contact element methods each provide valid ways to simulate bolts, rivets, and welds, their suitability depends on the project requirements.

## **5. Meshing strategies**

In industrial FEM practice, the primary priority, as mentioned previously, is to reduce cost and time. This objective also influences the meshing strategies ultimately adopted, which may differ from those most commonly recommended in academic

literature. The chosen mesh type, density, and generation methods are often determined not only by accuracy requirements but also by project deadlines, available computational resources, and engineer experience.

### **5.1 Meshing for 1D elements**

For 1D elements such as beams, meshing is generally straightforward. In most setups, a reasonable element size is selected based on the overall dimensions of the full model and the engineer's prior experience. Beam meshes rarely present quality issues, and the primary consideration is ensuring enough elements along with each member to capture global stiffness, deflection shapes, and modal behavior without oversizing the model.

### **5.2 Meshing for shell elements**

For shell elements, meshing is also relatively simple in most industrial scenarios. Almost exclusively, quadrilateral elements are generated, as they provide good numerical performance and stable results. Modern FEM programs are capable of meshing even complex shell structures using quadrilaterals without any significant difficulty.

The element size is again determined by the structural dimensions and the engineer's experience. While triangular shell elements were more common in earlier FEM work, especially for irregular geometries, they are now almost never the preferred choice since they can yield less accurate results.

### **5.3 Meshing for solid elements**

The meshing of solid elements varies significantly depending on the FEM software being used. Some programs can generate hex-dominant, high-quality meshes, which are generally more accurate and efficient. Others, however, rely mainly on tetrahedral elements for automated meshing of complex volumes.

The mesh size and refinement depend on the geometry of the component, structural requirements, and the engineer's judgment. In practice, it is common to find hybrid meshes, where some regions are meshed with tetrahedrons and others with hexahedrons. While this hybrid approach can solve meshing difficulties, it may introduce result deviations due to the different numerical properties of the element types—something that must be considered during post-processing and result interpretation.

### **5.4 Element order**

In most industrial applications, first-order elements (linear elements) are used almost exclusively. They are computationally faster, more stable in large models, and easier to mesh without generating distorted shapes. Second-order elements (quadratic elements) are typically reserved for more detailed local models or research-grade simulations, where capturing curvature and high-stress gradients is essential. While second-order elements can provide better accuracy per element, they also increase the number of degrees of freedom, solution times, and the risk of convergence issues in nonlinear simulations.

## 6. Material modeling

In industrial FEM applications, material models are selected with the same cost and time-driven priorities discussed earlier. As a result, linear elastic models with standard material property values are used almost exclusively. These values are typically taken from material handbooks, supplier datasheets, or built-in FEM software libraries.

In rare cases, companies may introduce their own experimentally derived material characteristics. This requires physical testing and data processing, which is an expensive process, and is therefore reserved for high-value projects, critical components, or situations where standard data is not sufficiently accurate.

Occasionally, nonlinear material models are introduced for specific components or detailed local studies. In such cases, bilinear material models are most common, as they are simpler to implement and require less input data while still capturing basic plastic behavior. Multilinear and constitutive models, which more accurately follow the full stress-strain curve, are typically reserved for research or highly specialized engineering work, as they demand extensive material data and significantly increase computational effort. As an example, **Figure 6** shows how different material models compare when plotted against the actual stress-strain curve. As the elastic limit is approached, the gap between the linear models and real behavior becomes clear. Even so, linear models are still the most commonly used in industries. The reason is simple: in most applications, plastic strains are not allowed, so the design process usually involves adjusting and iterating until the stresses remain within the elastic range – where linear models provide results as reliable as those of more complex nonlinear models.

In industrial environments, engineers will almost always try to formulate the problem as linear if this produces results that are “good enough” for the project requirements. Nonlinear material definitions are used only when strictly necessary, as they increase analysis time, complexity, and the risk of convergence issues.

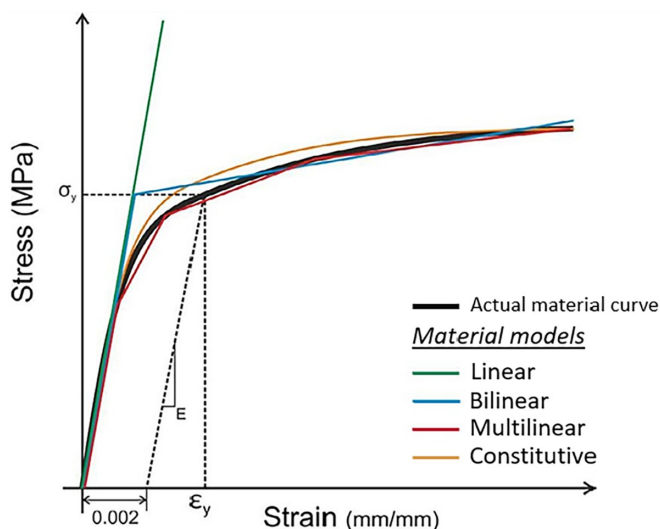


**Figure 6.**  
 Different material models over the material flow curve. Top-left: freight wagon frame. Adapted from [3].

## 7. Best practices summary for industrial FEM applications

The following table summarizes the most widely adopted best practices in industrial finite element modeling. It consolidates practical recommendations across all stages of FEM simulation from geometry preparation to material modeling

| Category                    | Best practice                                                                                                    | Industrial rationale                                                               |
|-----------------------------|------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------|
| Geometry preparation        | Remove non-critical details (logos, small fillets, tiny holes, textures).                                        | Reduces meshing complexity, prevents convergence issues, and speeds up simulation. |
|                             | Adapt the model to the chosen element type (mid-surface extraction for shells, keypoint/line for beams).         | Ensures that the geometry matches analysis requirements.                           |
|                             | Merge or contact shells appropriately, depending on accuracy needs.                                              | Prevents artificial stiffness and preserves realistic load transfer.               |
| Element selection           | Use beams for large tubular or frame structures.                                                                 | Fast to model and compute; sufficient for global stiffness/deflection analysis.    |
|                             | Use shells for thin-walled structures across most industries.                                                    | Best balance between accuracy, efficiency, and geometric fidelity.                 |
|                             | Use solids only for bulky parts or localized studies.                                                            | High accuracy for 3D stress fields, but computationally expensive.                 |
|                             | Use springs, dampers, and lumped masses with RBEs for non-structural behavior.                                   | Adds key behaviors without complex modeling.                                       |
| Connection modeling         | Use RBEs for simplified bolted, riveted, or welded joints.                                                       | Quick setup, scalable for large assemblies; post-process forces for code checks.   |
|                             | Consider MPCs when partial DOF coupling is needed.                                                               | Offers stiffness control without artificial mass.                                  |
|                             | Use contact elements with full DOF restrictions only if required.                                                | More realistic, but time-consuming and prone to convergence issues.                |
| Meshing strategies          | Beams: select mesh size based on model dimensions and experience.                                                | Prevents over-refinement while capturing the global response.                      |
|                             | Shells: use quadrilaterals as default; avoid triangles unless necessary.                                         | Maintains accuracy in bending-dominated problems.                                  |
|                             | Solids: prefer hex-dominant meshes; use tetrahedrons for complex volumes; accept hybrid meshes when unavoidable. | Balances accuracy with meshing ability.                                            |
|                             | Use first-order elements for most industrial cases; second-order elements only for detailed/local models.        | Reduces DOF count and solution time.                                               |
| Material modeling           | Use linear elastic models from trusted sources for most applications.                                            | Fast, stable, and adequate for many design needs.                                  |
|                             | Use bilinear models when basic plasticity needs to be captured.                                                  | Simple nonlinear approach without excessive complexity.                            |
|                             | Reserve multilinear or advanced nonlinear models for research or critical high-fidelity cases.                   | High data and computational costs.                                                 |
| General industrial approach | Always aim for the simplest model that meets accuracy requirements.                                              | Saves time, reduces costs, and ensures repeatable workflows.                       |
|                             | Validate key results against physical tests or simplified calculations.                                          | Ensures simulation credibility and compliance with standards.                      |

highlighting the rationale behind each approach. These guidelines are derived from real world engineering workflows, emphasizing efficiency, accuracy, and repeatability under industrial constraints.

## **8. Conclusion(s)**

The application of FEM in industrial environments is shaped as much by practical constraints as by technical capabilities. While modern FEM software and computational resources continue to evolve, providing increasingly sophisticated modeling tools, the driving factors in most industrial projects remain time, cost, and reliability. This reality fosters a set of pragmatic strategies that prioritize efficiency without compromising essential accuracy.

Geometry preparation is recognized as one of the most decisive stages in the simulation workflow. Models imported from CAD are rarely ready for direct analysis; therefore, defeaturing and model adaptation are standard practices aimed at removing non-critical details and tailoring the geometry to the intended simulation approach. Symmetry exploitation, although powerful, is seldom applied due to communication and validation challenges with clients and stakeholders.

Element-type selection is another cornerstone of industrial FEM. Beam and shell elements dominate due to their efficiency, flexibility, and ease of implementation. Solid elements, while theoretically versatile, are largely reserved for bulky parts or localized analyses because of their higher computational demands and sensitivity to mesh quality. Special-purpose elements, such as springs, dampers, mass elements with RBEs, and connection representations, enable engineers to include essential structural behaviors without unnecessary complexity.

The simulation of connections – bolted, riveted, or welded – follows the same philosophy. Engineers often favor simplified approaches using RBEs, MPCs, or beam-based connectors, supported by post-processing calculations to meet design standards. While more advanced methods, such as contact-based modeling, can provide higher fidelity, they are typically used only when required by the project scope or certification process.

Meshing strategies in industries reflect the same balance between accuracy and cost-efficiency. For beams, mesh sizing is straightforward. Shell models favour quadrilateral elements for stability and accuracy, while solid meshing varies with software capabilities, often resulting in hybrid meshes. First-order elements are the default choice for most industrial work, with second-order elements reserved for detailed studies where higher fidelity is essential.

Material modeling follows an equally pragmatic approach. Linear elastic properties from standard sources are used in the majority of cases. Nonlinear materials, whether bilinear or multilinear, are employed only when necessary, typically in detailed or research-level models, due to their higher computational requirements and data demands.

## **Author details**

Francisco Badea-Romero<sup>1\*</sup> and Jesús A. Perez-Fernandez<sup>2</sup>

1 Nebrija University, Madrid, Spain

2 Oviedo University, Oviedo, Spain

\*Address all correspondence to: Fbadea@nebrija.es

## **IntechOpen**

---

© 2025 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Liu WK, Li S, Park HS. Eighty years of the finite element method: Birth, evolution, and future. *Archives of Computational Methods in Engineering*. 2022;**29**:4431–4453. DOI: 10.1007/s11831-022-09740-9
- [2] Gohlke MJAE. Practical aspects of finite element simulation. In *Michigan. USA: Altair*; 2015
- [3] Alcalá E, Badea F, Martín Á, et al. Methodology for the accuracy improvement of FEM beam type T-junctions of buses and coaches structures. *International Journal of Automotive Technology*. 2013;**14**:817–827
- [4] Li C, Wu LL, Chai S, Zhao Y, Zhou YJ. Analysis of the influence of boundary conditions on the mechanical characteristics of arch bridge with composite arch. *IOP Conference Series: Earth and Environmental Science*. 2020;**446** (5):052035
- [5] Fernández JAP. Theoretical and experimental study of the fatigue behavior of large passenger transport vehicles under dynamic operating loads. *Diss Universidad Politécnica de Madrid*. 2017. DOI: 10.20868/UPM.thesis.47667
- [6] Zhang N-L, Gao C, Qiu G, Yang J-G, Wu B-J, Cheng -X-X. Non-stationary wind loading identification for large transmission tower based on dynamic finite-Element model updating. *Energies*. 2025;**18**:357. DOI: 10.3390/en18020357
- [7] Vega B, Pérez JÁ. Comparative analysis of fatigue strength of a freight wagon frame. *Weld World*. 2024;**68**:321–332. DOI: 10.1007/s40194-023-01577-5
- [8] Horváth P, Égert J. Stress analysis and weight reduction of a one-cylinder engine crankshaft. *Acta Technica Jaurinensis*. 2015;**8**(3):201–217. DOI: 10.14513/actatechjaur.v8.n3.377
- [9] Sun J, Yan S, Jiang Y, et al. Finite element analysis of the valgus knee joint of an obese child. *Bio Medical Engineering OnLine*. 2016;**15** (Suppl 2):158. DOI: 10.1186/s12938-016-0253-3
- [10] El Amli A, El Yousfi B, Soualhi A, Guillet F. A novel hybrid FEM–dynamic modeling approach for enhanced vibration diagnostics in a two-stage spur gearbox. *Energies*. 2025;**18**(9):2176



## Chapter 5

# Finite Element Method Applied to Heat Transfer in Food Processing

*Ana López-Palencia, Carlos Reales-Martínez and Fabián A. Ortega-Quintana*

### Abstract

This chapter presents finite element methods for solving engineering problems involving heat transfer mechanisms (conduction, convection, and radiation) in the context of food processing. It addresses problems governed by partial differential equations in blanching and cooling processes. The chapter begins with a general overview of the finite element method, followed by a focus on axisymmetric heat transfer problems. Cases with one and two subdomains in two dimensions are analyzed as an alternative approach to fully three-dimensional problems. Specifically, open-source finite element libraries (such as FEniCS, Gmsh, and ParaView) are employed to address issues in axisymmetric domains. Finally, examples of fruit blanching and cooling problems solved using the finite element method are provided.

**Keywords:** partial differential equations, *FEniCS*, heat transfer mechanism, blanching, cooling process, cucumber, carrot, avocado

### 1. Introduction

Transport phenomena (mass, energy, and momentum) form the essential foundation for the study and design of systems in process engineering. Their analysis is of great importance in areas such as chemical, mechanical, environmental, and food engineering, where physical processes are modeled and optimized using mathematical tools [1].

Within the field of food engineering, heat transfer (manifested by three mechanisms: conduction, convection, and radiation) stands out as a key process for ensuring thermal efficiency in preservation and processing operations. Temperature variations in food products directly affect their quality, organoleptic properties, nutritional profile, and shelf life. Therefore, the modeling and simulation of this phenomenon have been of special interest, particularly under realistic, non-ideal conditions and coupled with other mechanisms such as mass diffusion and fluid flow [2].

Thermal behavior in foods is closely linked to their geometry, composition, boundary conditions, and thermophysical properties. This requires the use of partial

differential equations (PDEs) to describe heat transport as a function of space and time. Given that analytical solutions to PDEs are generally restricted to ideal cases, numerical methods are used for their resolution. The finite element method (FEM) stands out for its ability to address irregular geometries and complex boundary conditions typical of food systems. By discretizing the domain into small elements, FEM allows obtaining internal temperature distributions, profiling critical superheating or cooling zones, and designing efficient thermal strategies for food processing. Although many studies have been limited to two-dimensional domains in transient problems, there is an opportunity to move towards three-dimensional models, especially in heterogeneous foods, and this consists of taking advantage of the axial symmetry of objects, which is the main objective of this chapter [3].

On the other hand, many physical problems, such as those related to heat transfer, are formulated as boundary value problems within three-dimensional domains, which implies a high computational burden due to the number of unknowns involved. An efficient solution consists of applying numerical methods that take advantage of the axial symmetry of the system to reduce the domain to a two-dimensional configuration, which is possible when the variables, parameters, and conditions of the problem are independent with respect to a spatial dimension. This approach is especially valid in axisymmetric systems, typical in thermal and mechanical phenomena, where the geometry and material properties are homogeneous (i.e., exhibit isotropy), allowing well-structured two-dimensional models to faithfully represent real three-dimensional behavior, minimizing computational resources without compromising the accuracy of the analysis.

This chapter presents specific examples of heat transfer, as well as the mathematical foundations for modeling heat transfer using PDE and numerical techniques such as FEM. Sections 2.1 and 2.2 explain important details of domains in cylindrical coordinates and provide a general description of the two-dimensional finite element method. Section 2.3 presents the numerical solution of heat transfer problems in compound and irregular axisymmetric domains, with varied boundary conditions and discretized in meshes that respect the subdomains of the geometries under study.

## **2. Finite element method**

Finite difference, finite element, and finite volume methods are often used to solve partial differential equations numerically because analytical solutions are generally limited to simple and idealized cases. The finite element method is used to solve partial differential equations that can be adapted to complex and irregular geometries and varied boundary conditions commonly found in food processing systems. The finite element method partitions the domain into smaller elements and approximates solutions to the equations, providing valuable information about temperature distributions, concentration profiles, and fluid flow patterns within the materials under study. The following briefly describes the preliminary and most relevant concepts when approaching the finite element method to solve heat transfer problems in food processes.

### **2.1 Preliminary concepts: Cylindrical coordinates and Sobolev spaces**

In this context, it is assumed that the domain  $\tilde{\Omega}$  is axisymmetric in cylindrical coordinates  $(r, \theta, z)$ , that is, it is symmetric with respect to the  $z$  axis, and the problem

parameters will be independent of the angular variable  $\theta$ . Under these hypotheses, the solution will be axisymmetric and its derivatives with respect to  $\theta$  will be zero. Therefore, it is sufficient to calculate the solutions in the two-dimensional domain  $\Omega = \{(r, z) : (r, 0, z) \in \tilde{\Omega}\}$ . In solving the problems presented in this chapter, the edge of the domain  $\partial\Omega$  intersects the axis of symmetry, a condition that must be considered in the mathematical and numerical formulation of the model.

Since, in the following methods, axisymmetric functions will be calculated, it is essential to express these functions in terms of the coordinates  $(r, \theta, z)$ . The unit vectors in cylindrical coordinates are denoted by  $e_r, e_\theta$  and  $e_z$ . So, a vector function is written as  $\mathbf{F} = (F_r, F_\theta, F_z)$  or  $\mathbf{F} = F_r e_r + F_\theta e_\theta + F_z e_z$ . Given a vector function  $\mathbf{F} = F_r(r, \theta, z)e_r + F_\theta(r, \theta, z)e_\theta + F_z(r, \theta, z)e_z$  and a scalar function  $f = f(r, \theta, z)$ , we have

$$\operatorname{div} \mathbf{F} = \frac{1}{r} \frac{\partial(rF_r)}{\partial r} + \frac{1}{r} \frac{\partial F_\theta}{\partial \theta} + \frac{\partial F_z}{\partial z} \quad (1)$$

$$\nabla f = \frac{\partial f}{\partial r} e_r + \frac{1}{r} \frac{\partial f}{\partial \theta} e_\theta + \frac{\partial f}{\partial z} e_z \quad (2)$$

In the following, a definition and a theorem necessary for the further development of the finite element methodology are stated [4]:

*Definition:* We introduce function spaces in  $\Omega$  that contain the traces of axisymmetric functions of *Sobolev* spaces defined on  $\tilde{\Omega}$ . Let  $L_r^2(\Omega)$  be the weighted *Lebesgue* space of all measurable functions  $A$  defined in  $\Omega$  such that

$$\|A\|_{L_r^2(\Omega)}^2 := \int_{\Omega} |A|^2 r dr dz < \infty \quad (3)$$

the corresponding inner product is defined by:

$$(A, Z)_{L_r^2(\Omega)} = \int_{\Omega} A Z r dr dz \quad (4)$$

The weighted *Sobolev* space  $H_r^k(\Omega)$  consists of all functions in  $L_r^2(\Omega)$  whose derivatives up to order  $k$  are also in  $L_r^2(\Omega)$ . We define the norm and the seminorm in the usual way, in particular:

$$\|A\|_{H_r^1(\Omega)}^2 := \int_{\Omega} (|\partial_r A|^2 + |\partial_z A|^2) r dr dz \quad (5)$$

Let  $L_{1/r}^2(\Omega)$  be the weighted *Lebesgue* space of all measurable functions  $A$  defined in  $\Omega$  such that:

$$\|A\|_{L_{1/r}^2(\Omega)}^2 := \int_{\Omega} \frac{|A|^2}{r} dr dz < \infty \quad (6)$$

The Hilbert space  $\tilde{H}_r^1(\Omega)$  is defined by  $\tilde{H}_r^1(\Omega) = \{A \in H_r^1(\Omega) : A \in L_{1/r}^2(\Omega)\}$  with the norm

$$\|A\|_{H_r(\Omega)}^{-1} := \left( \|A\|_{H_r^1(\Omega)}^1 + \|A\|_{L_{1/r}^2(\Omega)}^2 \right)^{1/2} \quad (7)$$

*Theorem: Integration by Parts.* For all  $w \in \tilde{H}_r^1(\Omega) \times H_r^1(\Omega)$  and  $\varphi \in H_r^1(\Omega)$ , we have

$$\int_{\Omega} \text{div } w \varphi r dr dz = - \int_{\Omega} w \nabla \varphi r dr dz + \int_{\partial\Omega} w \cdot \mathbf{n} \varphi r dS \quad (8)$$

The reader is referred to the proof of this theorem found in [5].

## 2.2 Two-dimensional finite element method

This section presents the main features of the finite element method in a two-dimensional configuration. It explains how, starting from the partial differential equation, a variational formulation and its discretization are obtained. This description will be presented using a stationary problem in cylindrical coordinates, followed by the application of some of the results obtained.

### 2.2.1 Variational formulation

The boundary value problem to be studied is the *Poisson* problem in cylindrical coordinates (Eq. (9)):

*Problem 1:* Find  $u(r, z)$  such that

$$\begin{cases} -\text{div}_{axi}(\nabla u) = \left( \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) \right) + \frac{\partial^2 u}{\partial z^2} = f, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega \end{cases} \quad (9)$$

The first step of the method is to provide a variational formulation. To do this, we multiply the differential equation in the value problem by the so-called test functions  $v(r, z)$  and integrate both sides over the entire domain  $\Omega$ :

$$\int_{\Omega} -\text{div}(\nabla u) v r dr dz = \int_{\Omega} f v r dr dz \quad (10)$$

Applying (Eq. (8)) to (Eq. (10)) we have:

$$\int_{\Omega} -\text{div}(\nabla u) v r dr dz = \int_{\Omega} \nabla u \nabla v r dr dz - \int_{\partial\Omega} \nabla u \cdot \mathbf{n} v r dS = \int_{\Omega} f v r dr dz \quad (11)$$

and since  $u = 0$  on the boundary of  $\Omega$ , the test function  $v$  will also assume that value, leading to (Eq. (12)):

$$\int_{\Omega} \nabla u \nabla v r dr dz = \int_{\Omega} f v r dr dz \quad (12)$$

Therefore, Problem 1 has been transformed into the following variational problem, which will be called Problem 2:

**Problem 2: variational formulation.** Find  $u \in \mathcal{V}_0 := \{v \in H_{r(\Omega)}^1 : u = 0 \text{ on } \partial\Omega\}$  such that  $a(u, v) = \int_{\Omega} f v r dr dz \forall v \in \mathcal{V}$ , where  $a(u, v) := \int_{\Omega} \nabla u \nabla v r dr dz$ .

As can be seen, compared to *Problem 1*, the unknown  $u$  is not required to be twice differentiable. The derivatives considered in *Problem 2* are weak derivatives instead of the classical ones of *Problem 1*. The pointwise equality on both sides has been replaced by an integral expression. For this reason, *Problem 2* is called a weak formulation and  $u$  a solution in the weak sense. However, it can be shown that under certain circumstances, a function  $u$  satisfying *Problem 2* also satisfies *Problem 1* classically [6].

### 2.2.2 Discrete formulation

Once the variational formulation is established, the next step in the finite element method is called *Galerkin* projection, with the goal of solving *Problem 2* on the computer. On the other hand, the infinite-dimensional space  $\mathcal{V}$  is difficult to manipulate, so  $\mathcal{V}$  is replaced by a finite-dimensional subspace  $\mathcal{V}_h$ .

For this purpose,  $\{T_h\}_{h>0}$  is considered a regular family of triangulations of  $\Omega$ , where  $h$  is the mesh size. Consider the space  $\mathcal{V}_h := \{v_h \in \mathcal{V} : v_h|_K \in \mathbb{P}_1 \forall K \in T_h\}$ , where  $\mathbb{P}_1$  are the real-valued linear functions in the coordinates  $r$  and  $z$ , i.e.,  $\mathbb{P}_1 := \{p(r, z) = c_0 + c_1 r + c_2 z : c_0, c_1, c_2 \in \mathbb{R}\}$ .

The finite element approximation of *Problem 2* must be found in the subspace  $\mathcal{V}_{h,0} := \{v_h \in \mathcal{V}_h : v_h = 0 \text{ on } \partial\Omega\}$ , therefore, *Problem 2* becomes the following *Problem 3*:

**Problem 3: Discrete variational formulation.** Find  $u_h \in \mathcal{V}_{h,0}$  such that  $a(u_h, v_h) = \int_{\Omega} f v_h r dr dz$  for all  $v_h \in \mathcal{V}_{h,0}$ .

Since  $\mathcal{V}_{h,0} \subset \mathcal{V}_0$  we have that  $a(u, v_h) = \int_{\Omega} f v_h r dr dz$  for all  $v_h \in \mathcal{V}_{h,0}$  and  $a(u_h, v_h) = \int_{\Omega} f v_h r dr dz$  for all  $v_h \in \mathcal{V}_{h,0}$ , from where we have that  $a(u - u_h, v_h) = 0$  for all  $v_h \in \mathcal{V}_{h,0}$ . This equation is known as *Galerkin* orthogonality [6].

### 2.2.3 Computational implementation

To construct a basis for the finite-dimensional space  $\mathcal{V}_h$  it is important to note that a function  $v \in \mathcal{V}_h$  is uniquely determined by its nodal values of  $\{v(N_j)\}_{j=1}^{n_p}$ .

Reciprocally, for every set of nodal values, there exists a unique  $v \in \mathcal{V}_h$  with those nodal values. For this reason, the nodal values will be the degrees of freedom and define a corresponding basis for  $\{\varphi_j\}_{j=1}^{n_p} \subset \mathcal{V}_h$ , such that

$$\varphi_j(N_i) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (13)$$

with  $i, j = 1, \dots, n_p$  where  $n_p$  corresponds to the number of nodes in the mesh.

As shown in **Figure 1**, each function  $\varphi_j$  is continuous and piecewise linear, with support only on the set of triangles that have a common vertex  $N_j$ . Now, using the basis functions, any function  $v \in \mathcal{V}_h$  can be written as (Eq. (14)):

$$v = \sum_{i=1}^{n_p} \alpha_i \varphi_i \quad (14)$$

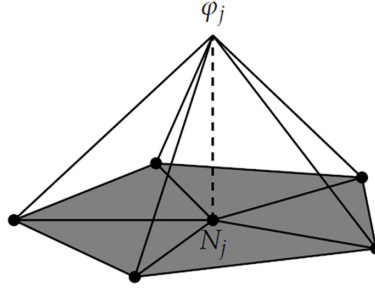


**Figure 1.**  
Hat functions in 2D.

where  $\alpha_i = v(N_i)$  with  $i = 1, \dots, n_p$  are the nodal values of  $v$ .

To compute the finite element approximation  $u_h$ , let  $\{\varphi_i\}_{i=1}^n$  be the basis for  $\mathcal{V}_h$  consisting of the hat functions associated with the  $N_i$  interior nodes within the mesh. No hat functions are needed at the mesh boundary, since the functions in  $\mathcal{V}_h$  vanish there. Now, using this basis, it is seen that the finite element method is equivalent to *Problem 4*:

*Problem 4:* Find  $u_h \in \mathcal{V}_{h,0}$  such that  $\int_{\Omega} \nabla u_h \nabla \varphi_i r dr dz = \int_{\Omega} f \varphi_i r dr dz \quad \forall i = 1, \dots, n$

The finite element approximation  $u_h$  belongs to the set  $\mathcal{V}_{h,0}$  so  $u_h$  is written as a linear combination of the basis functions, (Eq. (15)):

$$u_h(x) = \sum_{j=1}^{n_i} \xi_j \varphi_j(x) \quad (15)$$

with  $n_i$  and  $\xi_j$  ( $j = 1, \dots, n_i$ ), which are unknown to be calculated.

Therefore,

$$\int_{\Omega} f \varphi_i dx = \int_{\Omega} \nabla u_h \nabla \varphi_i dx = \int_{\Omega} \left( \sum_{j=1}^{n_i} \xi_j \nabla \varphi_j \right) \nabla \varphi_i dx = \sum_{j=1}^{n_i} \xi_j \int_{\Omega} \nabla \varphi_j \nabla \varphi_i dx \quad (16)$$

By introducing the notation  $A_{i,j} = \int_{\Omega} \nabla \varphi_j \nabla \varphi_i dx$  ( $i, j = 1, \dots, n_i$ ) and  $b_i = \int_{\Omega} f \varphi_i dx$  ( $i, j = 1, \dots, n_i$ ), we have  $b_i = \sum_{j=1}^{n-1} A_{i,j} \xi_j$  ( $i = 1, \dots, n-1$ ), which is a system of  $(n_i + 1) \times (n_i + 1)$  equations for the  $(n + 1)$  unknown coefficients  $\xi_j$ , with  $j = 0, 1, \dots, n$ . In matrix notation, it is written as (Eq. (17)):

$$A\xi = b \quad (17)$$

where the entries of the matrix  $(n_i + 1) \times (n_i + 1)$  of  $A$  are defined by the  $A_{i,j}$  and  $b_j$  defined above. This linear system, which is usually of high order, is solved with the help of specialized software (for example, the freely available software *Fenics*, used to obtain the solutions to the problems addressed in this chapter), hence obtaining the unknown variables  $\xi$  and in turn  $u_h$ .

#### 2.2.4 Finite element method in nonlinear problems.

Most of the physical, biological, and chemical processes occurring in nature are described by nonlinear mathematical models. Therefore, extensions of the finite element method to nonlinear equations are of particular interest. This section describes standard methods for addressing discretized nonlinear partial differential equations with finite elements, specifically the *Newton–Galerkin* method and its simplified variant, Picard, or fixed-point iteration [6].

-Picard iteration.

Consider the following problem: find  $u$  such that

$$\begin{cases} -\operatorname{div}(q(u)\nabla u) = f, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega \end{cases} \quad (18)$$

where  $q(u)$  is a nonlinear function. Picard iteration is an easy way to handle nonlinear partial differential equations: simply use a known prior solution in the nonlinear terms to make them linear in the unknown  $u$ . This strategy is also known as the method of successive substitutions. For our problem, a known prior solution in the coefficient  $q(u)$  is used.

Therefore, the problem in (Eq. (18)) is expressed as follows:

#### 2.3 Problem 5. Variational formulation.

Given  $u^0$ , find  $u^{k+1} \in \mathcal{V} := \{v \in H_r^1(\Omega) : u = 0 \text{ on } \partial\Omega\}$  such that

$$\int_{\Omega} q(u^k) \nabla u^{k+1} \nabla v \, dr \, dz = \int_{\Omega} f v \, dr \, dz \quad k = 0, 1, 2, \dots, \quad (19)$$

for all  $v \in \mathcal{V}$ .

Iterations stop when  $\epsilon \equiv \|u^{k+1} - u^k\| < \text{tol}$ , where  $\text{tol}$  is a small tolerance, say  $10^{-5}$ , or when the number of iterations exceeds some critical limit. The latter case will detect divergence from the method or unacceptably slow convergence. **Table 1** shows Picard's algorithm for a nonlinear problem.

---

##### Picard iteration for a nonlinear problem

---

- 1: Choose an initial solution  $u_h^0 \in \mathcal{V}_h$
- 2: for  $k = 0, 1, 2, \dots$ , do
- 3: Find the solution to

$$\int_{\Omega} q(u^k) \nabla u^{k+1} \nabla v \, dr \, dz = \int_{\Omega} f v \, dr \, dz$$

- 4: if  $|u^{k+1} - u^k| < \epsilon$  then
  - 5: Stop.
  - 6: end if
  - 7: end for
- 

**Table 1.**  
*Picard's algorithm.*

## 2.4 Newton–Galerkin method

Consider the problem described in (Eq. (18)). Given an approximation to the solution field,  $u^k$ , we seek a perturbation  $\delta u$  such that  $u^{k+1} = u^k + \delta u$  satisfies the nonlinear partial differential equation. However, the problem for  $\delta u$  is still nonlinear, so the idea is to assume that  $\delta u$  is small enough to linearize the problem with respect to  $\delta u$ . Inserting  $u^{k+1}$  into the partial differential equation and linearizing the  $q$  term as shown in (Eq. (20)):

$$q(u^{k+1}) = q(u^k) + q'(u^k)\delta u + \mathcal{O}_{((\delta u)^2)} \approx q(u^k) + q'(u^k)\delta u \quad (20)$$

eliminating nonlinear terms in  $\delta u$ , and rearranging, we obtain

$$\nabla \cdot (q(u^k) \nabla \delta u) + \nabla \cdot (q'(u^k) \delta u \nabla u^k) = f - \nabla \cdot (q(u^k) \nabla u^k) \quad (21)$$

The weak form of this partial differential equation is obtained by multiplying by a test function  $v$  and integrating over  $\Omega$  the piecewise second-order derivatives:

$$\int_{\Omega} (q(u^k) \nabla \delta u \cdot \nabla v + q'(u^k) \delta u \nabla u^k \cdot \nabla v) dx = \int_{\Omega} f v dx - \int_{\Omega} q(u^k) \nabla u^k \cdot \nabla v dx \quad (22)$$

Hence, the variational problem becomes: find  $\delta u \in V$  such that  $a(\delta u, v) = L(v)$  for all  $v \in \hat{V}$ , where:

$$\begin{cases} a(\delta u, v) = \int_{\Omega} (q(u^k) \nabla \delta u \cdot \nabla v + q'(u^k) \delta u \nabla u^k \cdot \nabla v) dx \\ L(v) = \int_{\Omega} f v dx - \int_{\Omega} q(u^k) \nabla u^k \cdot \nabla v dx \end{cases} \quad (23)$$

The function spaces  $V$  and  $\hat{V}$ , being continuous or discrete, are as in the linear Poisson problem shown above in (Eq. (9)).

If an initial guess is given, for example, the solution of the partial differential equation with  $q(u) = 1$ , the corresponding weak form  $a_0(u^0, v) = L_0(v)$  becomes:

$$a_0(u, v) = \int_{\Omega} \nabla u \cdot \nabla v dx, \quad L_0(v) = 0 \quad (24)$$

From here, we enter a cycle and solve  $a(\delta u, v) = L(v)$  for  $\delta u$  and so compute a new approximation  $u^{k+1} = u^k + \delta u$ . Identify that  $\delta u$  is a correction, so if  $u^0$  satisfies the prescribed *Dirichlet* conditions somewhere  $\Gamma_D$  of the boundary, we must require  $\delta u = 0$  on  $\Gamma_D$ . The *Newton–Galerkin* algorithm for solving a nonlinear problem is shown in **Table 2**.

| Newton–Galerkin iteration for a nonlinear problem                                                                                                                                               |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1: Choose an initial solution $u_h^0 \in \mathcal{V}_h$                                                                                                                                         |
| 2: for $k = 0, 1, 2, \dots$ , do                                                                                                                                                                |
| 3: Find the solution to                                                                                                                                                                         |
| $\int_{\Omega} \left( q(u^k) \nabla \delta u \cdot \nabla v + q'(u^k) \delta u \nabla u^k \cdot \nabla v \right) dx = \int_{\Omega} f v dx - \int_{\Omega} q(u^k) \nabla u^k \cdot \nabla v dx$ |
| 4: if $ u^{k+1} - u^k  < \epsilon$ then                                                                                                                                                         |
| 5: Stop.                                                                                                                                                                                        |
| 6: end if                                                                                                                                                                                       |
| 7: end for                                                                                                                                                                                      |

**Table 2.**  
 Newton–Galerkin algorithm.

## 2.5 Temperature distribution problems in food processing

### 2.5.1 Variational formulation in an axisymmetric domain

(Eq. (25)) is the general equation that represents energy transfer in a domain and in a transient state without chemical reaction in two dimensions [7]:

$$\begin{aligned}
 \rho C_p \partial_t T(t, r, z) &= \operatorname{div}(k \nabla T(t, r, z)) \text{ in } \Omega \\
 T(0, r, z) &= T_0(r, z) \text{ in } \Omega \\
 T &= T_1 \quad t \geq 0 \text{ on } \partial\Omega \\
 -k \frac{\partial T}{\partial n} &= h(T - T_\infty) \text{ at } t \geq 0, \text{ on } \partial\Omega
 \end{aligned}
 \tag{25}$$

The geometry and axisymmetric domain of this system are shown in **Figure 2**. (Eq. (25)) is valid in the  $\Omega$  domain,  $r$  and  $z$  are the geometric coordinates,  $\Gamma_1$  and  $\Gamma_0$

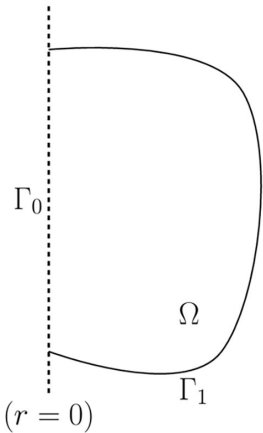


**Figure 2.**  
 2D geometry of an axisymmetric domain ( $\Omega$ ).

correspond to the convective interface and the interface where the temperature is prescribed, respectively,  $n = (nr, nz)$  is the unitary normal vector outside the surface. It should be noted that  $\Omega$  in this case corresponds to a surface and  $\partial\Omega = \Gamma_0 \cup \Gamma_1$  to a line that represents some edge of the domain, where the  $\Gamma_i \neq \emptyset$ . The same notation can be extended to three dimensions, where volumes and surfaces are involved. In this way, the mathematical problem with partial derivatives represented by the differential microscopic energy balance is posed.

To construct the variational formulation of (Eq. (25)), the differential equation of the system is multiplied by a test function  $v(r, z)$  and integrated over the entire domain  $\Omega$ :

$$\int_{\Omega} \rho C_p \partial_t T v r dr dz = \int_{\Omega} \text{div}(k \nabla T) v r dr dz \quad (26)$$

Applying (Eq. (8)), with  $w = k \nabla T$ , we obtain:

$$\int_{\Omega} \rho C_p \partial_t T v r dr dz = - \int_{\Omega} k \nabla T \nabla v r dr dz + \int_{\partial\Omega} k \nabla T \cdot n v r dS \quad (27)$$

Substituting the natural boundary conditions given in (Eq. (27)) into the boundary integral  $\partial\Omega$ , (Eq. (27)) becomes (Eq. (28)):

$$\int_{\Omega} \rho C_p \partial_t T v(r, z) r dr dz + \int_{\Gamma_1} h(T - T_{\infty}) v dS + \int_{\Omega} k \nabla T \cdot \nabla v r dr dz = 0 \quad (28)$$

It is observed that the essential boundary conditions (prescribed temperature) do not intervene in the variational formulation, hence integrating only over the boundary  $\Gamma_1$ , which establishes a flow condition at the interface (i.e., a derivative of the temperature). Therefore, the problem posed corresponds to finding function  $T$  that is  $T_1$  in  $\Gamma_0$  and that satisfies any admissible test function  $v$ :

$$\int_{\Omega} \rho C_p \partial_t T v r dr dz + \int_{\Omega} k \nabla T \cdot \nabla v r dr dz + \int_{\Gamma_1} h T v r dS = \int_{\Gamma_1} h T_{\infty} v dS \quad (29)$$

Applying the *Dirichlet* boundary conditions on  $\Gamma_1$ , taking the space  $\mathcal{V} = H_r^1(\Omega)$ , the weak formulation is described by the following problem:

---

PROBLEM A. Find  $T \in L^2(0, T, V)$  such that  $(\rho C_p \partial_t T, v)_{L_r^2(\Omega)} + a(T, v) = \ell(v)$   
 $\forall v \in \mathcal{V}$ , where:  $a(T, v) = \int_{\Omega} k \nabla T \cdot \nabla v r dr dz + \int_{\Gamma_1} h T v r dS$  and  $\ell(v) = \int_{\Gamma_3} h T_{\infty} v dS$

---

### 2.5.2 Variational formulation in an axisymmetric domain composed of two or more subdomains

In the previous development, a homogeneous material with a constant thermal coefficient  $k$  was assumed. However, in numerous scientific and engineering applications, it is necessary to consider systems composed of two or more materials, each with its own constant value of  $k$ , which is often omitted in conventional

analyses. Currently, this consideration has been incorporated as a more accurate strategy for estimating thermal solutions [8].

Below are problems in which determining factors are introduced that modify the classical formulation of the heat transport equation, such as the type of heat transfer mechanism, the presence of heat sources or losses, the geometry (regular or irregular), the differentiated compositions of the subdomains and their respective physicochemical properties, which will be properties of constant value or variable value.

*Case 1:* Let  $T(r, z)$  be the temperature distribution in a domain  $\Omega$  composed of the subdomains  $\Omega_1$  representing the inner region and  $\Omega_2$  representing the outer layer, **Figure 3** shows the relationship between the two subdomains.

Due to the discontinuous nature of the thermal conductivity parameter ( $k$ ), the differential equation is formulated as a split into subdomains  $\Omega_1$  and  $\Omega_2$ . Some continuity properties are provided along the interface, as shown in **Figure 3**. The domain boundaries are defined as follows:  $\partial\Omega_1 := \Gamma_0 \cup \Gamma_1 \cup \Sigma$  and  $\partial\Omega_2 := \Gamma_2 \cup \Sigma$  where  $\Gamma_0$  is the boundary that coincides with the rotation axis ( $r = 0$ ) and  $\Sigma$  is the interface or union of the two domains. The temperature  $T(r, z)$  is the solution to the following problem:

$$\begin{aligned}\rho_1 C_1 \partial_t T(t, r, z) &= \operatorname{div} \left( k_1 \nabla T(t, r, z) \right) \quad \text{in } \Omega_1 \\ \rho_2 C_2 \partial_t T(t, r, z) &= \operatorname{div} \left( k_2 \nabla T(t, r, z) \right) \quad \text{in } \Omega_2 \\ k_1 \mathbf{n} \cdot \nabla T|_{\Omega_1} &= k_2 \mathbf{n} \cdot \nabla T|_{\Omega_2} \quad \text{on } \Sigma \\ T|_{\Omega_1}(r, z) &= T|_{\Omega_2}(r, z) \quad \text{on } \Sigma \\ T(0, r, z) &= T_0(r, z) \quad \text{in } \Omega_1 \cup \Omega_2 \\ -k_1 \frac{\partial T}{\partial \mathbf{n}} \Big|_{\Gamma_1} &= h(T_\infty - T) \quad \text{on } \Gamma_1 \\ -k_2 \frac{\partial T}{\partial \mathbf{n}} \Big|_{\Gamma_2} &= h(T_\infty - T) \quad \text{on } \Gamma_2\end{aligned}\tag{30}$$

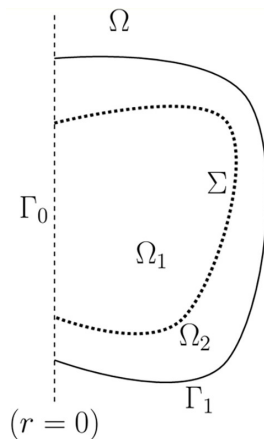


**Figure 3.**  
Domain ( $\Omega$ ) and its boundary ( $\partial\Omega$ ).

Multiplying the partial differential equations of expression (30) by a test function  $v$ , adding both and integrating over  $\Omega_1$  and  $\Omega_2$  separately, we obtain (Eq. (31)):

$$\begin{aligned} \int_{\Omega_1} \rho_1 C_1 \partial_t T v r d r d z + \int_{\Omega_2} \rho_2 C_2 \partial_t T v r d r d z - \int_{\Omega_1} \operatorname{div}(k_1 \nabla T) v r d r d z \\ - \int_{\Omega_2} \operatorname{div}(k_2 \nabla T) v r d r d z = 0 \end{aligned} \quad (31)$$

applying (Eq. (8)), we obtain (Eq. (32)):

$$\begin{aligned} \int_{\Omega_1} \rho_1 C_1 \partial_t T v r d r d z + \int_{\Omega_2} \rho_2 C_2 \partial_t T v r d r d z + \int_{\Omega_1} k_1 \nabla T \cdot \nabla v r d r d z + \int_{\Omega_2} k_2 \nabla T \cdot \nabla v r d r d z \\ - \int_{\partial \Omega_1} k_1 \nabla T \cdot \mathbf{n} v d S - \int_{\partial \Omega_2} k_2 \nabla T \cdot \mathbf{n} v d S = 0 \end{aligned} \quad (32)$$

In  $\Gamma_3 := \Gamma_1 \cup \Gamma_2$ , the natural boundary conditions are included, and together with the interface conditions, (Eq. (33)) is obtained:

$$\begin{aligned} \int_{\Omega_1} \rho_1 C_1 \partial_t T v r d r d z + \int_{\Omega_2} \rho_2 C_2 \partial_t T v r d r d z + \int_{\Omega_1} k_1 \nabla T \cdot \nabla v r d r d z + \int_{\Omega_2} k_2 \nabla T \cdot \nabla v r d r d z \\ + \int_{\Gamma_3} h T v d S = \int_{\Gamma_3} h T_\infty v d S \end{aligned} \quad (33)$$

Applying the *Dirichlet* boundary conditions on  $\Gamma_1$ , taking the space  $\mathcal{V} = H_r^1(\Omega)$ , the weak formulation is expressed as the following problem:

---

PROBLEM B. Find  $T \in \mathcal{V}$  such that  $m(\partial_t T, v) + a(T, v) = \ell(v) \quad \forall v \in \mathcal{V}$ , where

$$\begin{aligned} m(\partial_t T, v) &= \int_{\Omega_1} \rho_1 C_1 \partial_t T v r d r d z + \int_{\Omega_2} \rho_2 C_2 \partial_t T v r d r d z \\ a(T, v) &= \int_{\Omega_1} k_1 \nabla T \cdot \nabla v r d r d z + \int_{\Omega_2} k_2 \nabla T \cdot \nabla v r d r d z + \int_{\Gamma_3} h T v d S \\ \ell(v) &= \int_{\Gamma_3} h T_\infty v d S \end{aligned}$$


---

Note that “ $h$ ” refers to a subdomain parameter in the differential equation for calculating convective heat transfer. “ $h$ ” is used as a notation to refer to the mesh size within the discretization, so it is important to clarify that these terms are not related.

For the finite element approximation of PROBLEMS A and B, we proceed as follows:  $\{\mathcal{T}_h\}_{h>0}$  be a regular family of triangulations, where  $h$  is the size of the mesh that has a solution in the subspace  $\mathcal{V}_h := \{T_h \in C(\overline{\Omega}) \cap \mathcal{V} : T_h|_K \in \mathbb{P}_1(K) \quad \forall K \in \mathcal{T}_h\}$ , where  $\mathbb{P}_1$  are the complex-valued linear functions in the coordinates  $r$  and  $z$ ,  $\mathbb{P}_1 := \{p(r, z) = c_0 + c_1 r + c_2 z : c_0, c_1, c_2 \in \mathbb{R}\}$ , [9].

On the other hand, given a positive integer  $N$ , let  $\Delta t = t_{final}/N$ ,  $t_n = n\Delta t$  for  $0 \leq n \leq N$ , the implicit Euler method is considered, that is, the time derivative is approximated from (Eq. (34)):

$$\partial_t u = \frac{u(t^{n+1}) - u(t^n)}{\Delta t} = \frac{u^{n+1} - u^n}{\Delta t} \quad (34)$$

Therefore, PROBLEM B is reduced to Problem C:

---

PROBLEM C. Given  $T_h^0$ , find  $T_h^{n+1} \in \mathcal{V}_h$  such that

$$m(T_h^{n+1}, v) + \Delta t a(T_h^{n+1}, v) = \Delta t \ell(v) + m(T_h^n, v) \quad \forall v \in \mathcal{V}_h$$


---

Regarding the initial vector, there are a few different choices: one, and perhaps the simplest option is to let  $T_h^0$  be the nodal values of the interpolant  $\pi_h T_0$  of the initial data  $T_0$ ; another option is to let them be the nodal values of the  $L^2$  projection,  $P_h T_0$ , of  $T_0$ , but this is, of course, more computationally expensive. The discrete variational formulation expressed in PROBLEM C is widely used directly in computational codes. In this case, there is no need to impose boundary conditions.

*Case 2:* Let  $T(r, z)$  be the temperature distribution in a domain  $\Omega$  composed of subdomains  $\Omega_1$  representing the inner region and  $\Omega_2$  representing the outer layer. **Figure 3** above shows the relationship between the two subdomains. In this case, the physical properties depend on the temperature of the solid. The temperature  $T(r, z)$  is the solution to the following problem:

$$\begin{aligned} \rho_1(T) C_1(T) \partial_t T(t, r, z) &= \operatorname{div} \left( k_1(T) \nabla T(t, r, z) \right), \text{ in } \Omega_1 \\ \rho_2(T) C_2(T) \partial_t T(t, r, z) &= \operatorname{div} \left( k_2(T) \nabla T(t, r, z) \right), \text{ in } \Omega_2 \\ k_1(T) \mathbf{n} \cdot \nabla T|_{\Omega_1} &= k_2(T) \mathbf{n} \cdot \nabla T|_{\Omega_2}, \text{ on } \sum \\ T|_{\Omega_1}(t, r, z) &= T|_{\Omega_2}(t, r, z) \text{ on } \sum \\ T(0, r, z) &= T_0(r, z) \text{ in } \Omega_1 \cup \Omega_2 \\ -k(T) \frac{\partial T}{\partial \mathbf{n}} \Big|_{\Gamma} &= h(T_\infty - T) \text{ on } \Gamma \end{aligned} \quad (35)$$

After multiplying (Eq. (35)) by a test function and integrating by parts, the following variational problem is obtained:

---

Find  $T \in \mathcal{V}$  such that  $m(\partial_t T, v) + a(T, v) = \ell(v) \quad \forall v \in \mathcal{V}$ , where:

$$\begin{aligned} m(\partial_t T, v) &= \int_{\Omega_1} \rho_1(T) C_1(T) \partial_t T v r dr dz + \int_{\Omega_2} \rho_2(T) C_2(T) \partial_t T v r dr dz \\ a(T, v) &= \int_{\Omega_1} k_1(T) \nabla T \cdot \nabla v r dr dz + \int_{\Omega_2} k_2(T) \nabla T \cdot \nabla v r dr dz + \int_{\Gamma_3} h T v dS \\ (v) &= \int_{\Gamma_3} h T_\infty v dS \end{aligned}$$


---

This variational problem is reduced to solving PROBLEM C described as follows:

---

PROBLEM C: Given  $T_h^0$ , find  $T_h^{n+1} \in \mathcal{V}_h$  such that

$$m(T_h^{n+1}, v) + \Delta t a(T_h^{n+1}, v) = \Delta t \ell(v) + m(T_h^n, v) \quad \forall v \in \mathcal{V}_h$$


---

The formulation of this problem is nonlinear, so a strategy must be implemented in the computational code to address the nonlinearity. In this case, the Picard iteration was used, that is:

$$a(T_h^{n+1}, v) = \int_{\Omega_1} k_1(T_h^n) \nabla T_h^{n+1} \cdot \nabla v r dr dz + \int_{\Omega_2} k_2(T_h^n) \nabla T_h^{n+1} \cdot \nabla v r dr dz + \int_{\Gamma_3} h T_h^{n+1} v dS$$

*Case 3:* Let  $T(r, z)$  be the temperature distribution in a domain  $\Omega$  composed of the subdomains  $\Omega_1$  representing the inner region and  $\Omega_2$  representing the outer layer, as shown in **Figure 3**. For this case, there is a nonlinear function,  $Q_v(T)$ , in the equation of the domain or domains that depend on the temperature of the solid. The temperature  $T(r, z)$  is the solution to the following problem:

$$\begin{aligned} \rho_1 C_1 \partial_t T(t, r, z) &= \text{div} (k_1 \nabla T(t, r, z)) + Q_v(T) \text{ in } \Omega_1 \\ \rho_2 C_2 \partial_t T(t, r, z) &= \text{div} (k_2 \nabla T(t, r, z)) + Q_v(T) \text{ in } \Omega_2 \\ k_1 \mathbf{n} \cdot \nabla T|_{\Omega_1} &= k_2 \mathbf{n} \cdot \nabla T|_{\Omega_2} \text{ on } \sum \\ T|_{\Omega_1}(r, z) &= T|_{\Omega_2}(r, z) \text{ on } \sum \\ T(0, r, z) &= T_0(r, z) \text{ in } \Omega_1 \cup \Omega_2 \\ -k \frac{\partial T}{\partial \mathbf{n}} \Big|_{\Gamma} &= h(T_{\infty} - T) \text{ on } \Gamma \end{aligned} \quad (36)$$

The weak formulation of this problem, (Eq. (36)), is as follows:

---

Find  $T \in \mathcal{V}$  such that  $m(\partial_t T, v) + a(T, v) = \ell(v) \quad \forall v \in \mathcal{V}$ , where:

$$\begin{aligned} m(\partial_t T, v) &= \int_{\Omega_1} \rho_1(T) C_1(T) \partial_t T v r dr dz + \int_{\Omega_2} \rho_2(T) C_2(T) \partial_t T v r dr dz \\ a(T, v) &= \int_{\Omega_1} k_1 \nabla T \cdot \nabla v r dr dz + \int_{\Omega_2} k_2 \nabla T \cdot \nabla v r dr dz + \int_{\Omega_1} Q_v(T) v r dr dz + \int_{\Omega_2} Q_v(T) v r dr dz + \int_{\Gamma_3} h T v dS \\ \ell(v) &= \int_{\Gamma_3} h T_{\infty} v dS \end{aligned}$$


---

The time variable is discretized as in the previous cases, which generates the following problem:

---

Given  $T_h^0$  find  $T_h^{n+1} \in \mathcal{V}_h$  such that:

$$m(T_h^{n+1}, v) + \Delta t a(T_h^{n+1}, v) = \Delta t \ell(v) + m(T_h^n, v) \quad \forall v \in \mathcal{V}_h$$


---

Notice that the term

$$\begin{aligned} a(T_h^{n+1}, v) &= \int_{\Omega_1} k_1 \nabla T_h^{n+1} \cdot \nabla v r dr dz + \int_{\Omega_2} k_2 \nabla T_h^{n+1} \cdot \nabla v r dr dz + \int_{\Omega_1} Q_v(T_h^{n+1}) v r dr dz \\ &\quad + \int_{\Omega_2} Q_v(T_h^{n+1}) v r dr dz + \int_{\Gamma_3} h T_h^{n+1} v dS \end{aligned}$$

is nonlinear due to the  $Q_v$  function. This requires finding a solution using one of the numerical methods described above. In the example presented in the next section to illustrate this case, the *Newton–Galerkin* method is employed, as it is efficiently implemented in the *FEniCS* library used in this work.

To conclude this section, it is clarified that there are heat transfer processes where cases 2 and 3 are combined, that is, the physical properties of the solid depend on the temperature, and there is a function in the subdomains that is also dependent on the temperature. These cases are left to the reader as a complementary exercise to this chapter.

### 2.5.3 Application of the finite element method in food processing

*Case 1: Cucumber Blanching Process:* Blanching is the process in which foods (usually fruits and vegetables) are briefly immersed in high-temperature water

(70–100°C) and then rapidly cooled to stop cooking. The technique deactivates enzymes that affect the flavor, color, and texture of products, allowing them to be stored in optimal conditions. This first example considers a thermal process of blanching a cucumber in hot water, whose information was taken from [9]. The mathematical model of heat transfer expressed in (Eq. (30)) was used. The dimensions of the cucumber taken for the resolution of the mathematical model were 0.20 m high and 0.036 m in diameter at the widest part of the fruit. The initial temperature of the cucumber is 25°C, the temperature of the blanching medium is 80°C, and the value of the convective heat transfer coefficient was 643.6W/m<sup>2</sup>°C. The physical parameters, shown in **Table 3**, are considered constant over time.

The resolution of (Eq. (30)) by the finite element method is based on solving PROBLEM B, since it is considered that the cucumber is composed of two parts or domains (mesocarp and endocarp) with constant but different physical parameters. PROBLEM B is reduced to PROBLEM C, and this is solved with the computational code found in <https://github.com/careales/BookHeatTransferExamples>

**Figure 4** shows the temperature dynamics at three different points on the cucumber: one in the center, one in the mesocarp, and one on the surface in contact with the blanching fluid, corresponding to the coordinates (0,0.1), (0.009,0.1), and (0.018,0.1), respectively. This behavior is typical of the heating processes of most foods: first, there is a slow increase in temperature; second, the temperature increases rapidly until, finally, the food reaches a steady state with the heating fluid. Furthermore, the three heating curves for each simulated point do not overlap, indicating a strong dependence of heat transfer on spatial coordinates.

The above is also visualized in a differential color diagram, such as **Figure 5**, which shows the temperature distribution of the cucumber at two different blanching times, at 100 s and 1,000 s.

*Case 2: Carrot blanching process:* The heat transfer equation presented in the cucumber blanching example represents a linear problem, since the values of the thermal properties are constant and therefore independent of temperature. In addition, the boundary conditions are also considered linear. An example of carrot blanching is presented below, whose heat transfer equation, (Eq. (35)), is posed for two domains (see **Figure 3**). The thermal properties are dependent on the temperature and the chemical composition of the food (water, carbohydrates, proteins, fats, ash, and fiber), that is, the problem is non-linear. The case was solved using the carrot dimensions shown in **Table 4**.

Furthermore, the initial food temperature and the blanching fluid temperature were 20.1°C and 80°C, respectively. The convective heat transfer coefficient was 433.6 W/m<sup>2</sup>°C. The physical properties of the carrot were calculated from the Choi and Okos Equations [10], considering the proximal composition of the two carrot

|                         | Density ( $\rho$ ), kgm <sup>-3</sup> | Thermal conductivity ( $\kappa$ ),<br>W m <sup>-1</sup> K <sup>-1</sup> | Specific heat ( $C_p$ ),<br>J kg <sup>-1</sup> K <sup>-1</sup> |
|-------------------------|---------------------------------------|-------------------------------------------------------------------------|----------------------------------------------------------------|
| Mesocarp ( $\Omega_1$ ) | 959                                   | 0.61                                                                    | 4130                                                           |
| Endocarp ( $\Omega_2$ ) | 959                                   | 0.57                                                                    | 4120                                                           |

**Table 3.**  
*Physical parameters of the cucumber.*

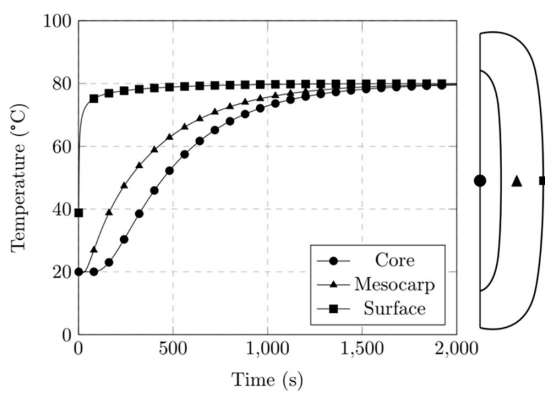


**Figure 4.**  
*Temperature profiles in blanched cucumber.*

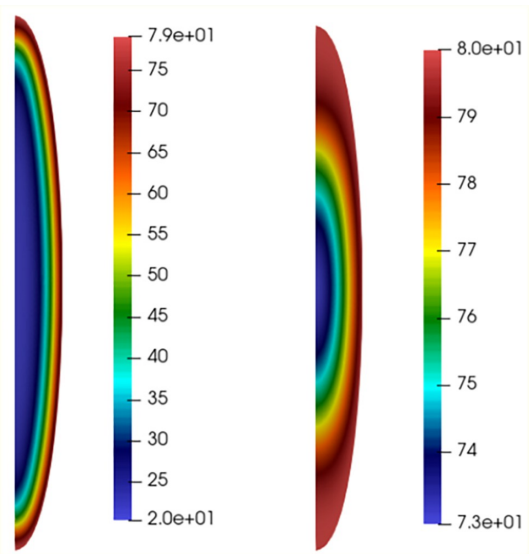


**Figure 5.**  
*Temperature distributions (°C) at 100 s and 1000 s in the cucumber blanching process.*

| Dimension                     | Value, m |
|-------------------------------|----------|
| Cylinder length               | 0.0538   |
| Largest diameter in mesocarp  | 0.0436   |
| Largest diameter in endocarp  | 0.0255   |
| Smallest diameter in mesocarp | 0.0380   |
| Smallest diameter in endocarp | 0.0200   |

**Table 4.**  
*Dimensions for the carrot.*

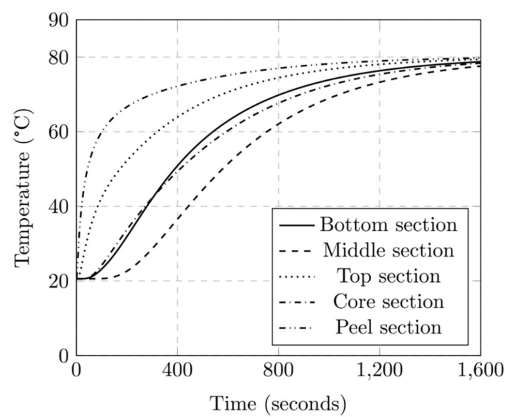


**Figure 6.**  
*Temperature profiles in blanched carrots.*

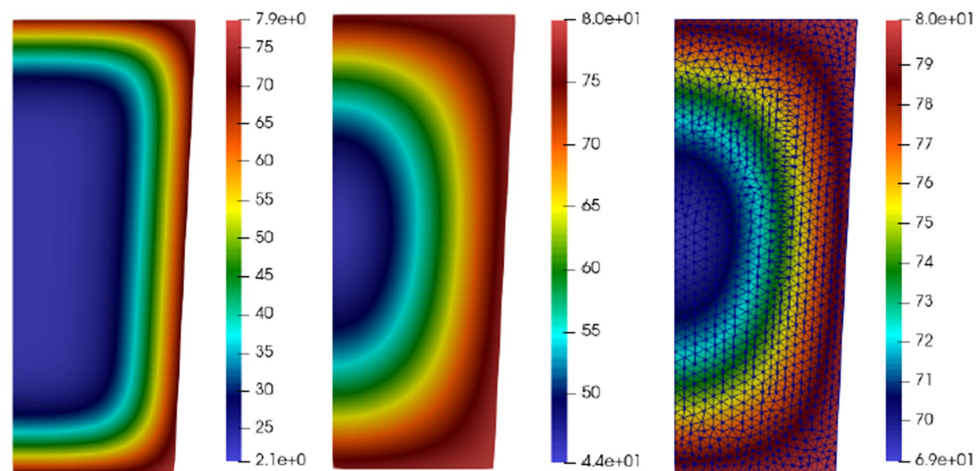


**Figure 7.**  
*Temperature distributions (°C) at  $t = 100$  s,  $500$  s, and  $1000$  s.*

layers reported in [11]. This is solved with the computational code found in <https://github.com/careales/BookHeatTransferExamples>

**Figure 6** shows the temperature dynamics at different points in the carrot geometry studied: bottom point (0, 0.01), middle point (0, 0.027), top point (0, 0.05), heart point (0.0105, 0.027), and rind point (0.019, 0.027). This behavior is like what occurs in the blanching of a cucumber, which was already analyzed in the previous example..

On the other hand, **Figures 7** and **8** show the temperature distribution at different processing time values (100 s, 500 s, and 1000 s). It is observed that there are no abrupt temperature changes at the cortex/core interface, which is achieved thanks to the correct connection of the nodes at the interface between one domain and another.

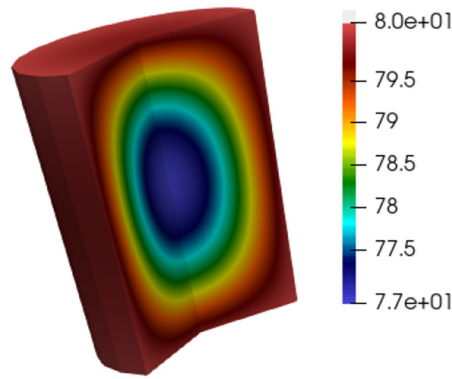


**Figure 8.**  
Temperature distribution ( $^{\circ}\text{C}$ ) in three-dimensional domain,  $t = 1,500 \text{ s}$ .

|                     | Density ( $\rho$ ),<br>$\text{kgm}^{-3}$ | Thermal conductivity ( $\kappa$ ),<br>$\text{W m}^{-1}\text{K}^{-1}$ | Specific heat ( $C_p$ ),<br>$\text{J kg}^{-1} \text{K}^{-1}$ |
|---------------------|------------------------------------------|----------------------------------------------------------------------|--------------------------------------------------------------|
| Pulp ( $\Omega_1$ ) | 977                                      | 0.425                                                                | 3460                                                         |
| Seed ( $\Omega_2$ ) | 1172                                     | 0.454                                                                | 2765                                                         |

**Table 5.**  
Physical parameters of avocado.

*Case 3: Avocado Cooling Process:* In this last example of this section, an avocado undergoing air precooling is considered [12]. Heat generation resulting from respiration is accounted for, while radiative losses at the fruit surface are considered negligible. As in the previous examples, an axisymmetric model suitable for this type of fruit (Eq. (36)) is used, and a finite element method is also used to approximate the temperature distribution. Heat generation due to respiration is estimated from (Eq. (37)):

$$Q_v(T) = 513 - 142T + 12T^2 - 0.22T^3 (\text{Wm}^{-3}) \quad (37)$$

The specific heat and thermal conductivities of the two components of the domain are estimated using the equations proposed for the avocado fruit based on the water and oil contents, shown in **Table 5**.

This problem is of constant physical parameters, however, the heat of respiration is dependent on the temperature and, in addition, the degree of the (Eq. (37)) makes it classified as a nonlinear problem, as indicated for the (Eq. (36)). The initial temperature is  $30^{\circ}\text{C}$ , the temperature of the cold air is  $10^{\circ}\text{C}$  and the value of the convective coefficient of heat transfer was  $643.6 \text{ W/m}^2\text{C}$ . This is solved with the computational code found in <https://github.com/careales/BookHeatTransferExamples>

**Figures 9, 10, and 11** show the temperature profiles at five (5) points of the avocado: Point 1 (0,0.06), Point 2 (0.02,0.06), Point 3 (0.03,0.06), Point 4 (0.01,0.1), and Point 5 (0.02,0.03). This behavior is typical in the cooling processes of most foods: first, there is a slow decrease in temperature; second, the

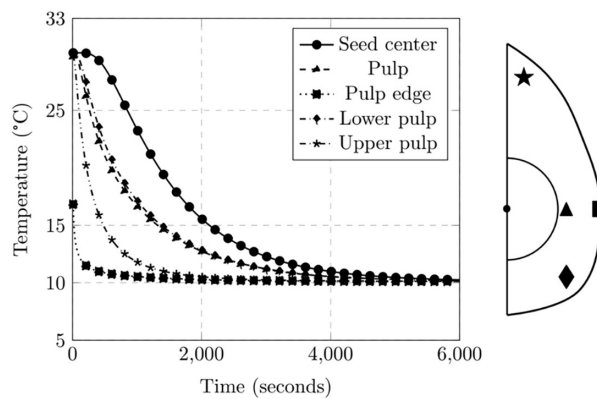


**Figure 9.**  
*Temperature profiles in avocado.*

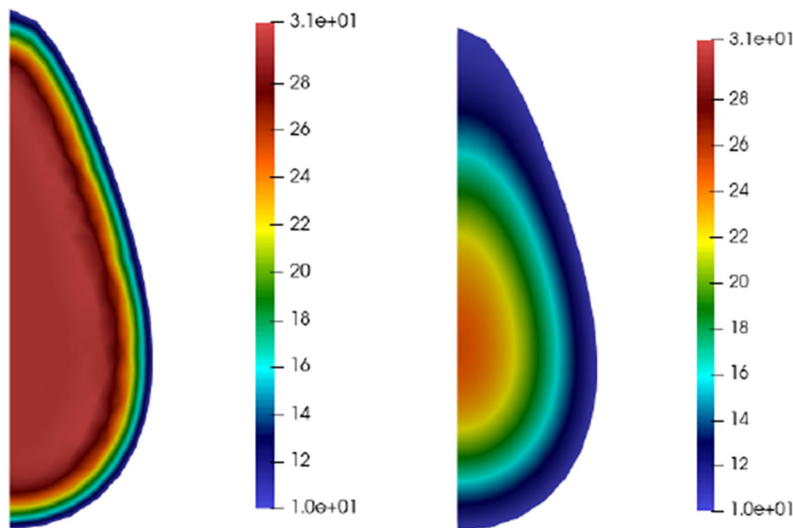


**Figure 10.**  
*Temperature distributions (°C) at  $t = 100$  s and  $1000$  s.*

temperature decreases rapidly until the food reaches a steady state with the cooling fluid.

### 3. Conclusion

The application of the finite element method (FEM) to heat transfer in food processing represents a significant advance in the optimization of thermal processes. From a mathematical perspective, the use of partial differential equations (PDEs) and their solution using numerical techniques such as FEM allows for the approach to complex problems with irregular geometries and varied boundary conditions. This approach not only facilitates the derivation

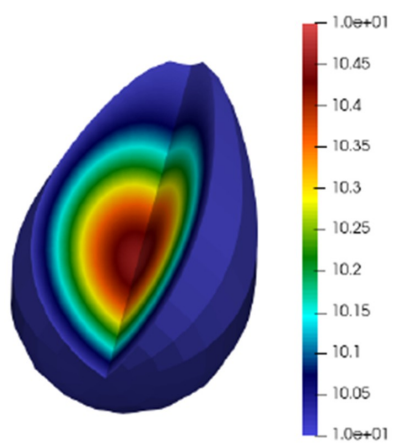


**Figure 11.**  
*Temperature distributions (°C) in three-dimensional domain,  $t = 5,000$  s.*

of internal temperature distributions but also allows for the identification of critical zones of overheating or cooling, which is essential for designing efficient thermal strategies.

In the field of food engineering, heat transfer is a fundamental process that directly affects the quality, organoleptic properties, nutritional profile, and shelf life of food products. The ability of FEM to model and simulate this phenomenon under realistic and non-ideal conditions is crucial for ensuring thermal efficiency in preservation and processing operations. The examples presented, such as blanching cucumber and carrot, as well as cooling avocado, demonstrate the applicability of FEM in the food industry. In closing, the integration of advanced mathematical tools with specific food engineering knowledge allows for improved understanding and control of thermal processes, therefore contributing to innovation and optimization in the food industry.

### Nomenclature

|            |                                                                  |
|------------|------------------------------------------------------------------|
| $C_p$      | specific heat of the material ( $\text{Jkg}^{-1}\text{K}^{-1}$ ) |
| $h$        | convective heat transfer coefficient                             |
| $t$        | time (s)                                                         |
| $T$        | temperature (°C)                                                 |
| $T_\infty$ | temperature of the heating fluid (°C)                            |
| $T_0$      | initial temperature (°C)                                         |

### Greek letters

|        |                                                         |
|--------|---------------------------------------------------------|
| $\rho$ | Density ( $\text{kgm}^{-3}$ )                           |
| $k$    | Thermal conductivity ( $\text{Wm}^{-1}\text{°C}^{-1}$ ) |

## Author details

Ana López-Palencia, Carlos Reales-Martínez and Fabián A. Ortega-Quintana\*  
Universidad de Córdoba, Montería, Colombia

\*Address all correspondence to: [fortega@correo.unicordoba.edu.co](mailto:fortega@correo.unicordoba.edu.co)

## IntechOpen

---

© 2025 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Bird R, Stewart W, Lightfoot E. Transport Phenomena. 1<sup>st</sup> ed. New York: Wiley; 2006. p. 928
- [2] Castro A, Mayorga E, Moreno F. Mint: Mathematical modelling of convective drying of feijoa (*Acca sellowiana* Berg) slices. Journal of Food Engineering. 2019;252:44–52. DOI: 10.1016/j.jfoodeng.2019.02.007.
- [3] Puri V, Anantheswaran R. 1993. Mint: The finite-element method in food processing: A review. Journal of Food Engineering. 19(3):247–274. DOI: 10.1016/0260-8774(93)90046-M.
- [4] Adams R, Fournier J. Pure and Applied Mathematics. 2<sup>nd</sup> ed. Amsterdam: Elsevier/Academic Press; 2003. p. 305
- [5] Gopalakrishnan J, Pasciak J. Mint: The convergence of V-cycle multigrid algorithms for axisymmetric Laplace and Maxwell equations. Mathematics of Computation. 2006;75:1697–1719
- [6] Khennane A. Introduction to finite element analysis using Matlab and Abaqus. 1<sup>st</sup> ed. Boca Raton: CRC Press; 2013. p. 453
- [7] Çengel Y, Ghajar A. Heat and Mass Transfer. Fundamentals & Applications. 5<sup>th</sup> ed. New York: McGraw Hill; 2015. p. 968
- [8] Yeo Y. Chemical Engineer Computation with Matlab. 2nd. Boca Raton: CRC Press; 2020. p. 847. DOI: 10.1201/9781003090601.
- [9] Fasina O, Fleming H. Mint: Heat transfer characteristics of cucumbers during blanching. Journal of Food Engineering. 2001;47:203–210. DOI: 10.1016/S0260-8774(00)00117-5.
- [10] Ortega-Quintana F, Pérez-Sierra O, Tarrá-Lozano L, López-Acosta E. Mint: Modelado Matemático de la Transferencia de Calor del Proceso de Escaldado de Zanahoria (*Daucus carota* L.). Información Tecnológica. 2017;28(6):3–10
- [11] Valente M, Chambarel A, Cordonnier J, Pumborios M. Mint: Finite element modelling of heat transfer in avocados. International Agrophysics. 1996;10:123–129
- [12] Choi Y, Okos M. Mint: Effects of Temperature and Composition on the Thermal Properties of Foods. Journal of Food Engineering and Process Applications. 1986;1:93–101

## Chapter 6

# Finite Element Analysis for Micromixing Applications

*Imran Shah, Babar Hayat and Shijun Ji*

### Abstract

Micromixers play an important role in microfluidic systems by enhancing fluid mixing at the microscale. This fluid mixing is vital for applications in biomedical, thermal, chemical, and energy harvesting systems. Finite element analysis (FEA) provides efficient numerical methods and simulation tools for simulating fluid flow and mass transfer within various micromixers. This chapter focuses on the application of FEA for micromixing, including experimental and numerical studies on Y-shaped split-and-recombine (SAR) micromixers, Cantor fractal-based micromixers, and obstacle-based micromixers. The comparative critical discussions highlight the optimal parametric design, mixing index, and performance evaluation for different real-time applications. An index above 80% is acceptable as per micromixing criteria. All proposed designs studied and included in this chapter have a mixing index in the range of 80–99% for Reynolds numbers of 10–100. These designs, with efficient mixing capability, can be implemented for real-time applications.

**Keywords:** micromixers, mixing index, finite element analysis (FEA), computational fluid dynamics (CFD), microfluidics, and mixing efficiency

### 1. Introduction

Micromixers are the main components in microfluidic systems, which enable efficient fluid mixing at very small length scales called the microscale. At the microscale, fluid flows through small channels, and the flow is laminar [1]. On the macroscale, fluid flow is turbulent, and fluids are mixed due to convection and advection without the use of a micromixer. In microscale mixing, molecular diffusion and convection do not occur. Introducing convection and producing mixing effects in microchannels requires certain geometrical changes for passive mixing, while in active mixing [2], some external energy sources must be connected to produce a disturbance in the flow and create a mixing effect. There are advantages to passive micromixers, as they do not require external energy, but their geometry is difficult to fabricate. In the active case, fast mixing is possible, but the integration of the energy component is challenging. COMSOL is a finite element analysis (FEA)-based tool used for FEA, solving the Navier–Stokes and convection–diffusion equations to provide a robust numerical solution for micromixer flow. The concentration diluted

module is used for mixing species of different concentration fluids; usually, 0 and 1 are used, and ultimately, 0.5 represents the mixed condition.

## **2. Background on micromixing and FEA**

Micromixing can be categorized into two main types: active and passive [3]. Active micromixers work using external sources of energy, such as electricity and pressure. Similarly, acoustic fields and magnetic fields are used for active mixing. On the other hand, passive micromixers do not require any external source of energy and utilize geometric features to enhance fluid mixing. Y-shaped split-and-recombine (SAR) structures, Y-shaped Cantor fractals, and M-shaped obstacle designs are widely investigated and focused on in this chapter for their better mixing capability. The FEA tool, COMSOL, enables the simulation of these designs by modeling fluid dynamics and mass transport at microscale dimensions. The performance of these proposed designs is compared using line plots and surface plots to measure their mixing capability for the selection of a better design.

## **3. Finite element modeling methodology**

### **3.1 Modeling steps**

The finite element method (FEM) is used to model and solve the micromixer problem using stiffness matrices and other techniques. However, if matrices exceed 12 by 12, it becomes too difficult and time-consuming to solve the problem [4]. For these larger matrices, FEA tools like COMSOL are widely used for micromixer simulations to discretize complex geometries and solve them accurately in less time. The FEM techniques are coded and utilized for FEA analysis in the software [5]. The software is used as a black box by providing input, defining physics, and obtaining the solution as an output accordingly. The simulation methodology typically involves the following steps:

- a) Geometry making of micromixer designs (Y-shaped, fractal, or obstacle-based).
- b) Meshing/Discretization of the microchannel domain/geometry into finite elements.
- c) Applying the boundary conditions, such as inlet velocity, outlet pressure, and no-slip walls, within the physics.
- d) Solution of fluid flow by Navier–Stokes and species by transport (convection–diffusion) equations after computation and running the simulation.
- e) The post-processing is conducted to evaluate the concentration and velocity profiles distribution, along with the mixing index calculation.

The details of all these steps are represented in **Figure 1**.

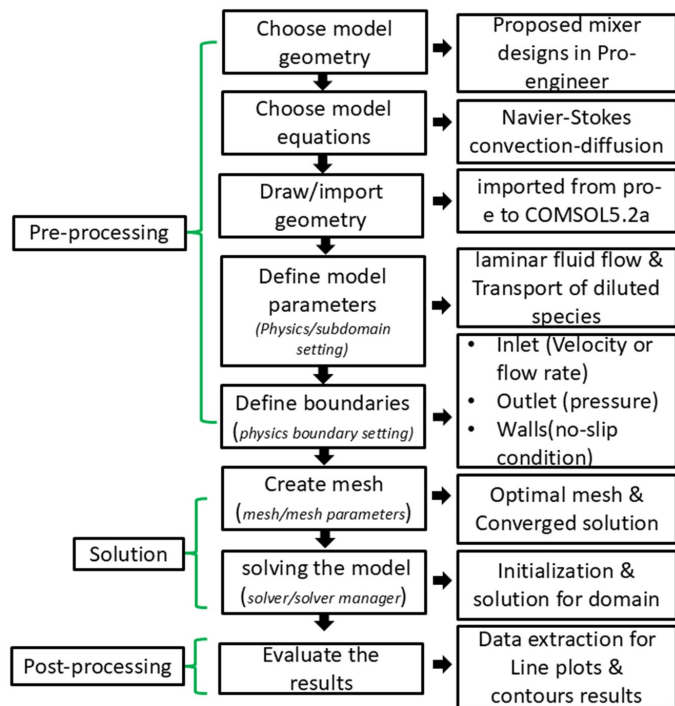


**Figure 1.**  
Represents the flow diagram for simulation processes in different steps.

### 3.2 Physics for micromixers

The following equations of Navier-Stokes and the concentration of diluted species are utilized with boundary conditions of inlet velocity and outlet pressure. No no-slip boundary condition is used on the walls of the channel.

Steady equation:

$$\rho \nabla \cdot \vec{V} = 0 \quad (1)$$

Navier-Stokes equation:

$$\rho(\mathbf{V} \cdot \nabla) \mathbf{V} = \nabla \cdot [-P\mathbf{I} + \mu(\nabla \mathbf{V} + (\nabla \mathbf{V})^T)] + \mathbf{F} \quad (2)$$

Convection-diffusion equation

$$\nabla \cdot (-D \nabla C_i) \mathbf{C} + \mathbf{V} \cdot \nabla C_i = R \quad (3)$$

$$\mathbf{V} = \mathbf{V}_0, \text{ at inlets} \quad (4)$$

$$\mathbf{V} = 0, \text{ at walls} \quad (5)$$

$$P = 0, \text{ at outlet} \quad (6)$$

where  $\mathbf{V}$  is velocity,  $P$  is pressure,  $\mathbf{F}$  is the body force, and  $R$  is the source term.

### 3.3 Mixing index calculation

The mixing index calculation formula is given by Equations 7–9 as presented below [6]. A mixing index value of 0 represents an unmixed situation, while a value of 1 represents fully mixed conditions.

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (c_i - c)^2} \quad (7)$$

$$\sigma_{max} = \sqrt{C(1 - C)} \quad (8)$$

$$M = 1 - \sqrt{\frac{\sigma^2}{\sigma_{max}^2}} \quad (9)$$

where M is the mixing index,  $\sigma$  is the standard deviation,  $\sigma_{max}$  is the maximum standard deviation, C is the concentration average value of the standard input of 0 and 1 at two inlets, and  $C_i$  represents any value on the cross-sectional plane as shown in **Figure 2**. Usually, 400  $C_i$  values are enough for accuracy.

### 3.4 Mesh Independence test

The numerical simulation conducted through the FEM, based on COMSOL software, requires accuracy in results. To ensure this accuracy, the mesh is initiated with an initial normal or fine mesh, and then it is refined to finer and extremely fine mesh settings by default within the COMSOL mesh settings. Usually, an experienced user begins with a slightly fine mesh that leads to quick convergence and stability of results. The simple meaning of this test is that if the size of the mesh is changed, there will be no significant effect on the results, and this point is referred to as mesh independence. An optimal mesh: “A mesh with low computational cost and enough accuracy is called an optimal mesh.” This optimal mesh is selected based on the results of the mesh independence test, and then all simulations are performed using this mesh exclusively. For example, **Figure 3** shows a meshing sample with

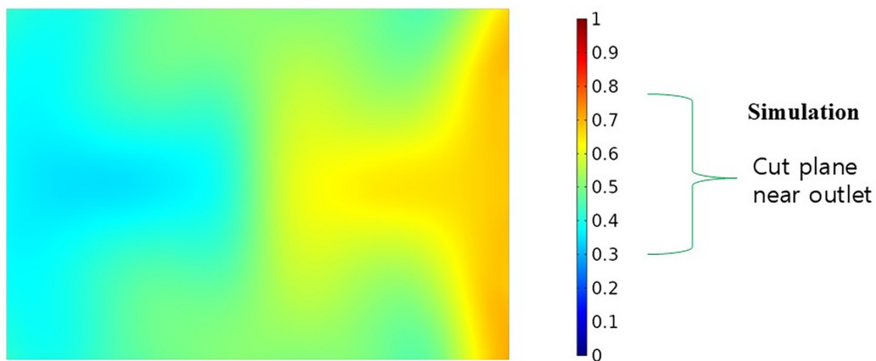


**Figure 2.**  
Represents the cut plan result on the cross-sectional plan of the micromixer channel for mixing index calculation.

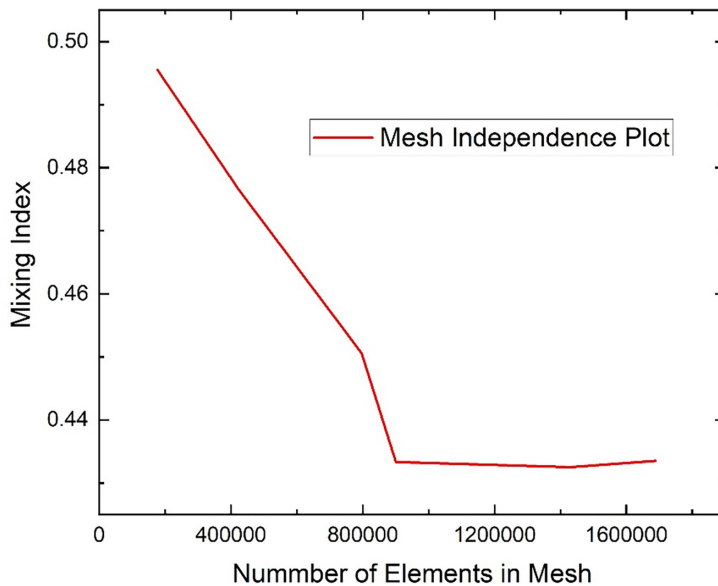


**Figure 3.**  
 Shows the mesh-independent test result for the number of elements versus mixing index.

several line plot results. The number of elements is changed from 177,359 to 1,688,555, and the results for the last three numbers of elements are very close. Therefore, the mesh with 900,150 elements is selected as the optimal mesh to save computation, and at this point, the accuracy of the results is sufficient or remains consistent.

Similarly, the optimal mesh detail can be added to research from mesh statistics, just like the number of elements, as discussed above. These may be vertices, faces, etc. An example of an optimal mesh with a magnified view of different portions for easy visualization is presented in **Figure 4**.

### 3.5 Validation of results

The FEA numerical model needs to be validated against experimental work or simulation. If the results are within a 10% error, they are considered authentic and acceptable. There are three kinds of solutions for any problem: analytical, simulation, and experimental. The strongest solution is analytical, but it is limited to regular geometry and areas where mathematical models are available. Experimentation involves fabricating and testing the material or structure through destructive or non-destructive testing. However, random, systematic, and apparatus errors may lead to inaccurate solutions. Proper care is required, and in some cases, repeated testing is necessary, which increases material costs, fabrication costs, labor, and equipment expenses. These factors make the solution very expensive, and experimental work requires cross-verification to ensure the authenticity of the results. The third approach is numerical simulation. In this case, many assumptions are made during the modeling phase, such as the flow being viscous and incompressible, and the absence of temperature effects on density, among others. Additionally, truncation errors and the neglect of higher-order derivatives in numerical schemes within the software can lead to predictive closure but not fully accurate results. These results

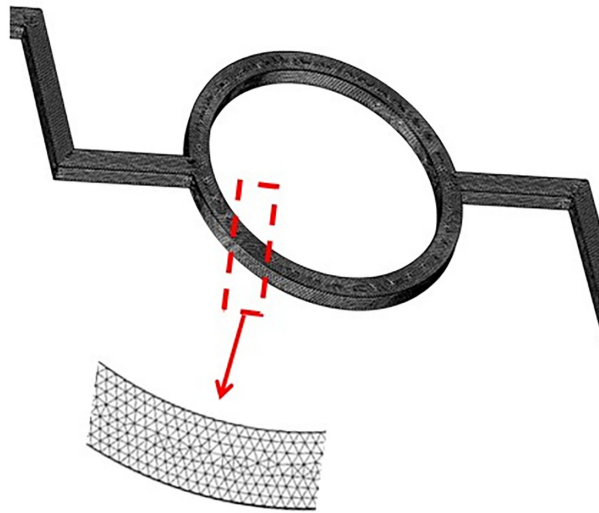


**Figure 4.**  
*Optimal mesh sample for the numerical simulation model with a magnified view.*

can sometimes be misleading, so they also need to be cross-verified through analytical, experimental, or literature-based data. For example, in the case of an analytical model tested for simple regular geometry, once it works correctly, it can be applied to a complex case for which an analytical solution is not possible. The benefit of simulation lies in its requirement for skill and a high-specification PC, which can provide predictive results and even optimization before practical testing, thereby saving significant costs and time.

### *3.5.1 Why is validation needed*

In order to validate the model for the case, there is no opportunity to verify results from analytical solutions or even experimental work. For example, in an analytical case, mathematical equations like turbulence advanced modeling are not yet available, or the geometry is so small that it cannot be fabricated, making testing through experiments impossible. In such a specific case, the only option is to validate with an already published paper in the literature, preferably experimental, but it can also be compared and validated with a previously published numerical paper as evidence that the model is working correctly. During validation, the same conditions used in the experiment or simulation are applied, such as creating the same size geometry and applying physics, etc., as per the technical problem. The results are then compared using surface color plots or line plots and presented in the research article. The same model, after verification, is applied to the proposed case to claim new research results for a new article as their own contribution to the existing literature.

### *3.5.2 Types of validation*

The two types of validation for FEA numerical analysis results are qualitative and quantitative.

### 3.5.2.1 Qualitative validation

The result of the current model for validation is compared with the surface, arrow plot, and the pattern of color distribution, and the legend shows the same scale magnitude of values for the desired parameter. Then, these compared results of the current model with the literature are referred to as qualitative validation. An example of qualitative validation is shown in **Figure 5**.

### 3.5.2.2 Quantitative validation

The result of the current model for validation is compared with any of the line plots, and the trends remain the same, along with the % error being below 10%, which means the model is validated. Thus, these compared results of the current model with the literature line plot are called quantitative validation. An example of quantitative validation is presented below in **Figure 6**.

## 4. Case studies of different micromixer novel designs

### 4.1 Passive micromixers

#### 4.1.1 Y-shaped split-and-recombine (SAR) micromixers

The experimental and numerical investigations of Y-shaped SAR micromixers demonstrate their effectiveness in enhancing the mixing effect by dividing and recombining the fluid flow channel in the micromixer design [9]. The results indicate that, upon completion of these studies, increasing the number of mixing units significantly improves the mixing index due to the availability of a more enhanced interfacial area between fluid layers. Additionally, an increase in the Reynolds number raises the speed of flow, and due to flow distribution and rotation, mixing is

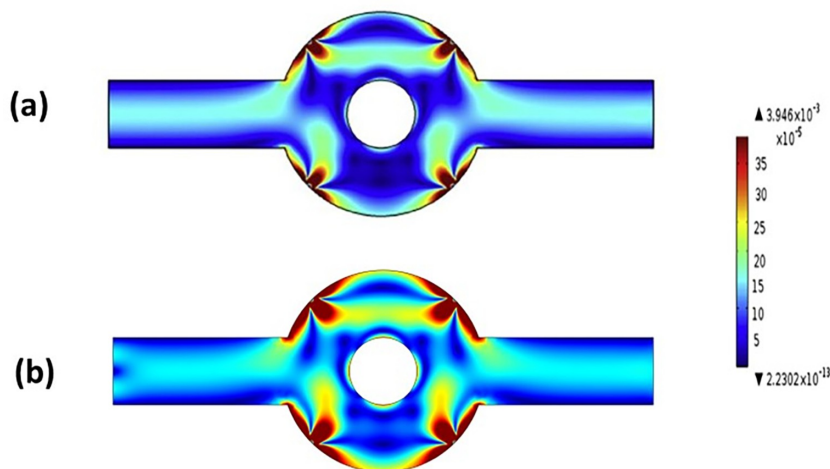


**Figure 5.**  
 Represents the qualitative validation of the (a) numerical model [7] and (b) current model results.

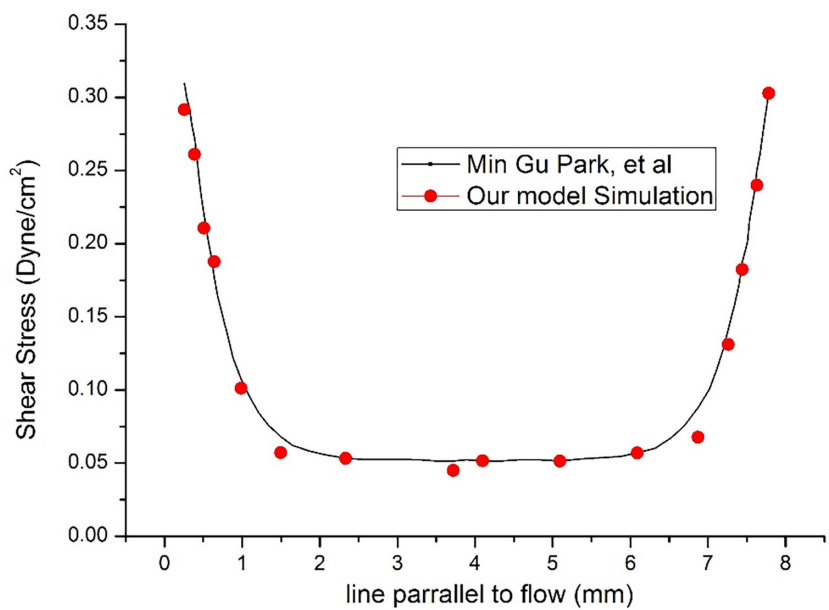


**Figure 6.**  
*Quantitative validation by comparing results with the published paper by Min Gu Park et al. [8].*

enhanced as swirling and Dean flow effects are produced inside the channel. These two factors influence the mixing and improve the mixing index. In the initial two to three units, the mixing is at a low level; however, as the fluid progresses through additional mixing units, there is a significant increase in the mixing index as convection dominates over diffusion. The experimental setup for testing practical micromixers is shown in **Figure 7**. The fabricated chip and its magnified view during testing are shown in **Figure 8**.

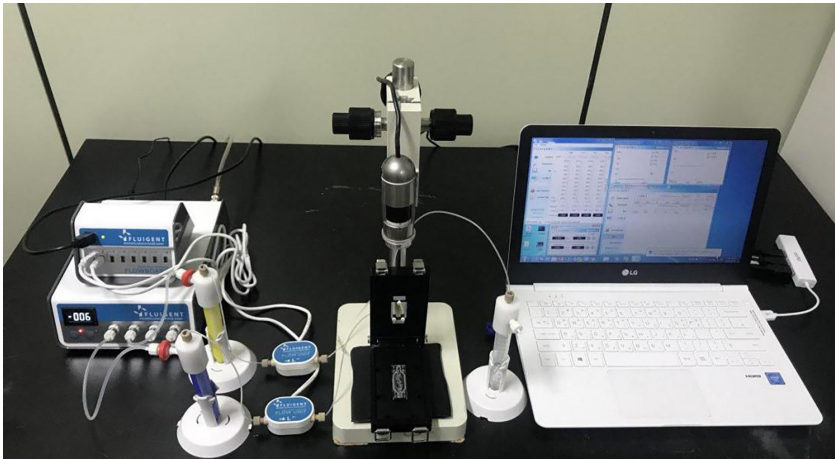


**Figure 7.**  
*Experimental setup for micromixer testing.*

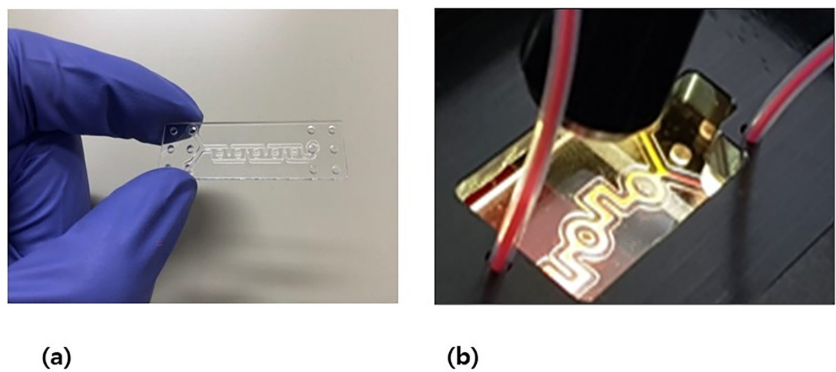


**Figure 8.**  
(a) Fabricated chip and (b) magnified view of fluid flow testing during experiment.

The simulation of a straight-shaped, SAR-based proposed micromixer is presented below in **Figure 9**.

The straight channel shows very low-quality mixing, as there is only diffusion in the middle of the channel. Meanwhile, the proposed SAR demonstrates a good quality of mixing, achieving 93% due to both diffusion and convection occurring.

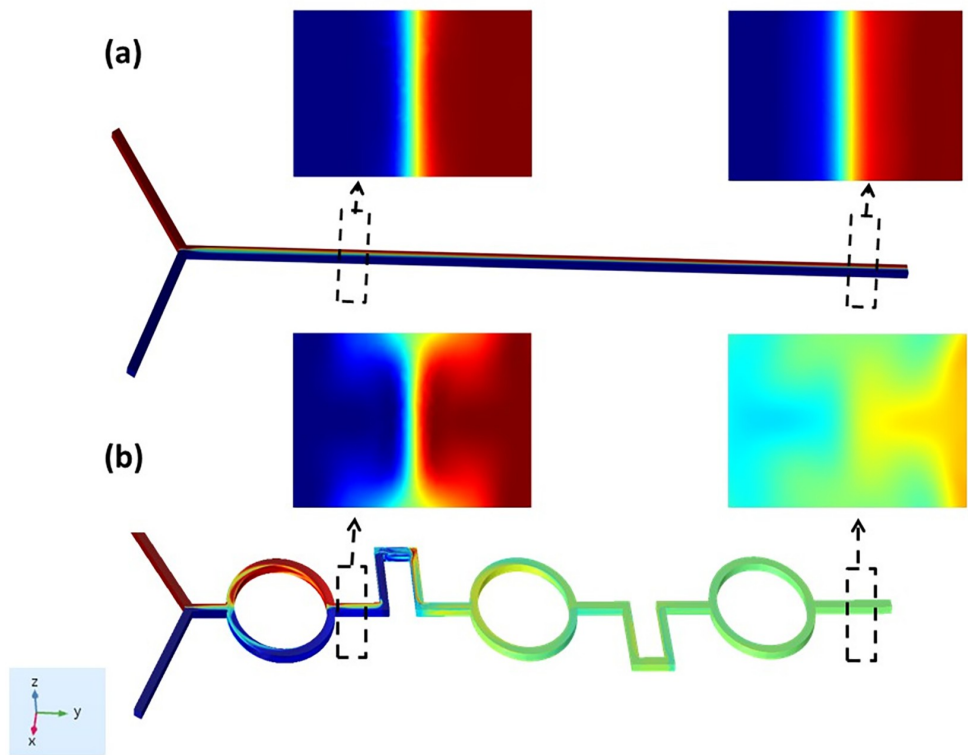


**Figure 9.**  
Comparison of mixing quality between the straight and SAR-based micromixer.

This is attributed to the mixing units and the division of the channel, which produce rotation in the fluid flow.

#### *4.1.2 Cantor fractal-based micromixers*

The Cantor fractal-based micromixer designs have been investigated numerically in COMSOL 6.3 and then tested experimentally. These structures increase the interfacial area, producing chaotic advection that results in superior mixing performance. The finite element COMSOL models confirm the scalability of fractal-based micromixers for good mixing and were implemented in real-time for biodiesel synthesis applications [10]. The comparison of experimental and simulation designs of micromixers for Cantor fractal-based ideas is presented in **Figure 10**.

The mixing index is low in the lowest case of the Reynolds number, as there is only diffusion and no convection occurring [11].

#### *4.1.3 Obstacle-based micromixers*

Obstacle-based micromixers introduce M-shaped barriers within the channel of the mixer design. When fluid passes through these shapes, it creates secondary flows, rotation, swirling effects, and vortices that enhance the mixing performance. Numerical studies have been carried out using FEA on non-aligned M-shaped micromixers. The results show mixing improvement with a good level of mixing index [12]. These micromixers were used for biomedical applications in laminar flow regions, particularly at low and moderate Reynolds numbers. The M-shaped micromixer's best performance design is presented in **Figure 11**.

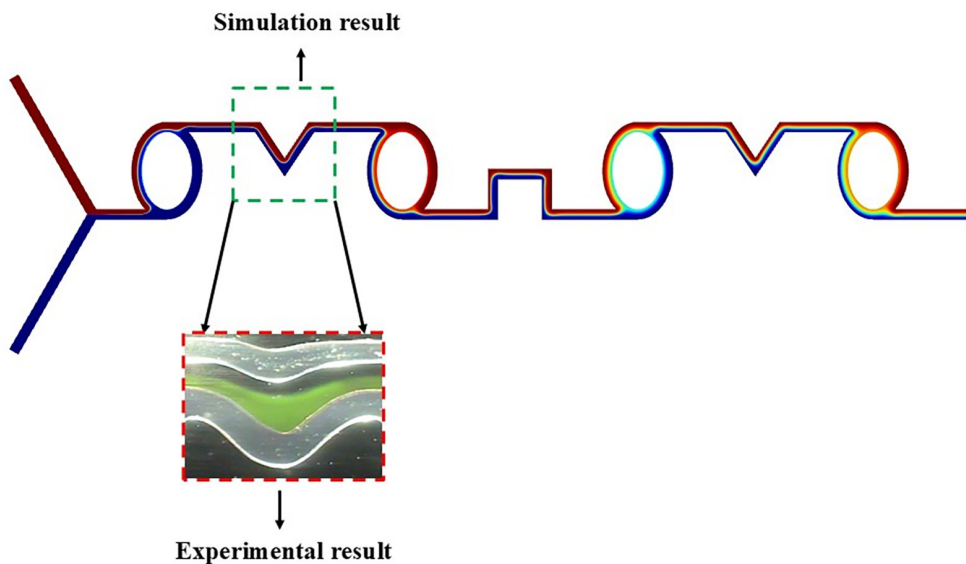


**Figure 10.** Cantor fractal-based micromixer simulation and experimental comparison showing agreement at the lowest Reynolds number of 0.5.

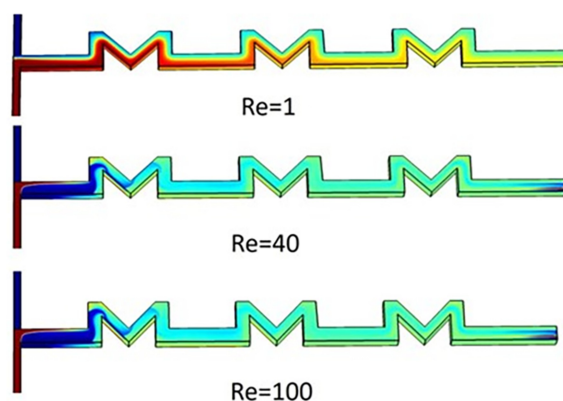


**Figure 11.**  
 Represents the M-shaped micromixer with the best performance of 95%.

## 4.2 Active micromixers

### 4.2.1 Parametric optimization of an active micromixer using the Taguchi method

The Taguchi method, using Minitab software, has been employed to optimize the active micromixers by studying the effect on mixing efficiency with different sets of input parameters for voltage, frequency, and velocity. Through COMSOL simulations and numerical studies, the optimal parameter sets were selected, which minimized the pressure drop and maximized the mixing index performance.

**Figure 12** represents the active micromixer's geometry and results at an optimal set of input parameters selected based on the Taguchi method, along with a complete analysis and testing of nine cases. These nine cases were evaluated at different values of frequency, velocity, and voltage. The result is an obtained frequency of 8 Hz, velocity of  $0.1 \text{ mm s}^{-1}$ , and voltage of 0.5 V [13].

### 4.2.2 Active mixing by surface acoustic wave

Active mixing is possible through external energy applied, producing a turbulence effect at low Reynolds numbers.

In **Figure 13**, a schematic diagram representing the mixing idea through the surface acoustic wave (SAW) approach is shown. The voltage is applied as an input to the interdigital transducer (IDT).

The voltage produces a vibration effect inside the piezoelectric substrate, and a transverse wave is generated on the surface of the substrate, called a SAW, which interacts with the fluid droplet. Due to the SAW interaction, a mixing effect is produced inside the droplet. This approach produces fast mixing with a high mixing index [14]. The mixing effect produced inside the droplet at different passages of time is presented in **Figure 14**. The results show that mixing is produced inside the droplet while the SAW interacts, and the piezoelectric substrate generates it in the y-direction.

**Figure 15** shows the cut plan result in the middle of the droplet for checking the mixing effect inside it. These surface contours represent the y-direction displacement or deformation produced due to the SAW production in the substrate.

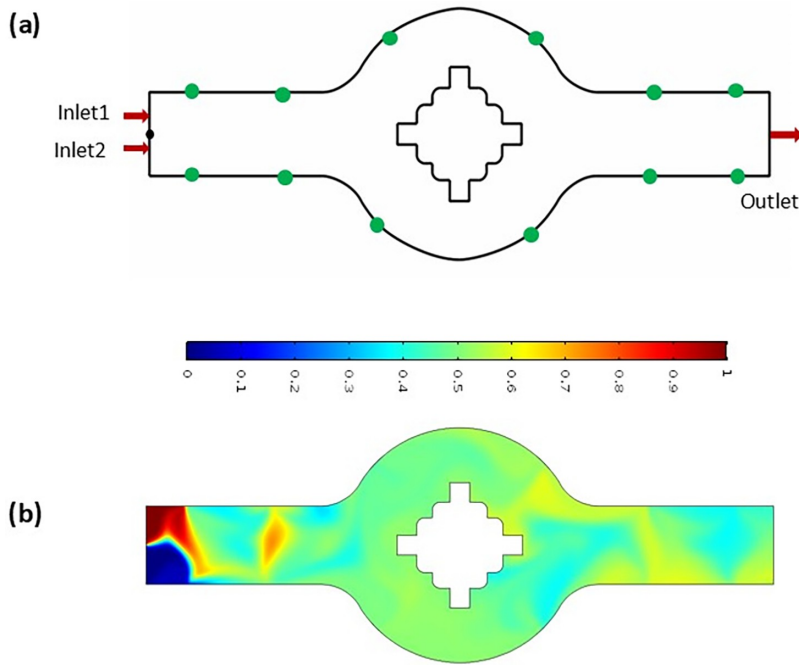


**Figure 12.** Represents the active micromixer: (a) geometry with a green dot indicating electrode positions and (b) the optimal set for better mixing effect is frequency 8 hz, velocity 0.1 mm/s, and voltage 0.5 V, obtained based on the Taguchi optimization method.

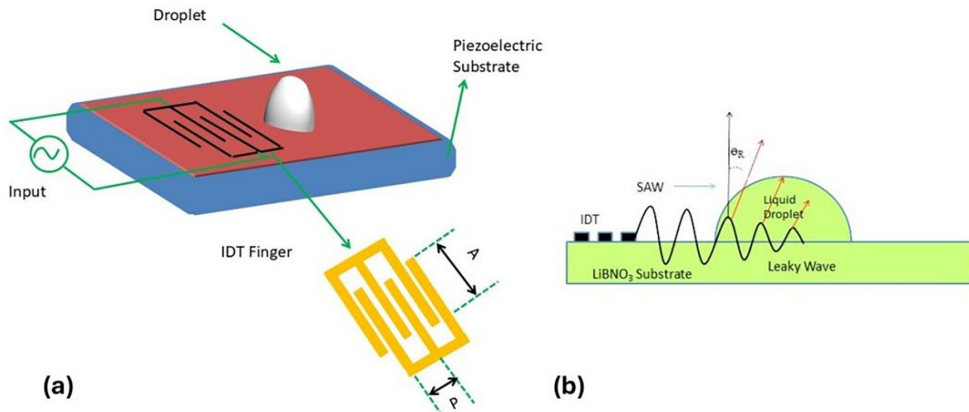


**Figure 13.** Represents the active micromixing through surface acoustic wave (SAW).

The color legend, with a blue minimum value and a red maximum value, indicates that the displacement effect varies at different locations, and hence, the mixing effect also changes in the same manner. The y-direction displacement is directly proportional to the mixing effect. This means that the higher the displacement, the greater the ability to produce a mixing effect, and vice versa. Mixing through the SAW concept is very fast and can produce homogeneous mixing, which can be implemented in real-time applications for point-of-care (POC) devices.

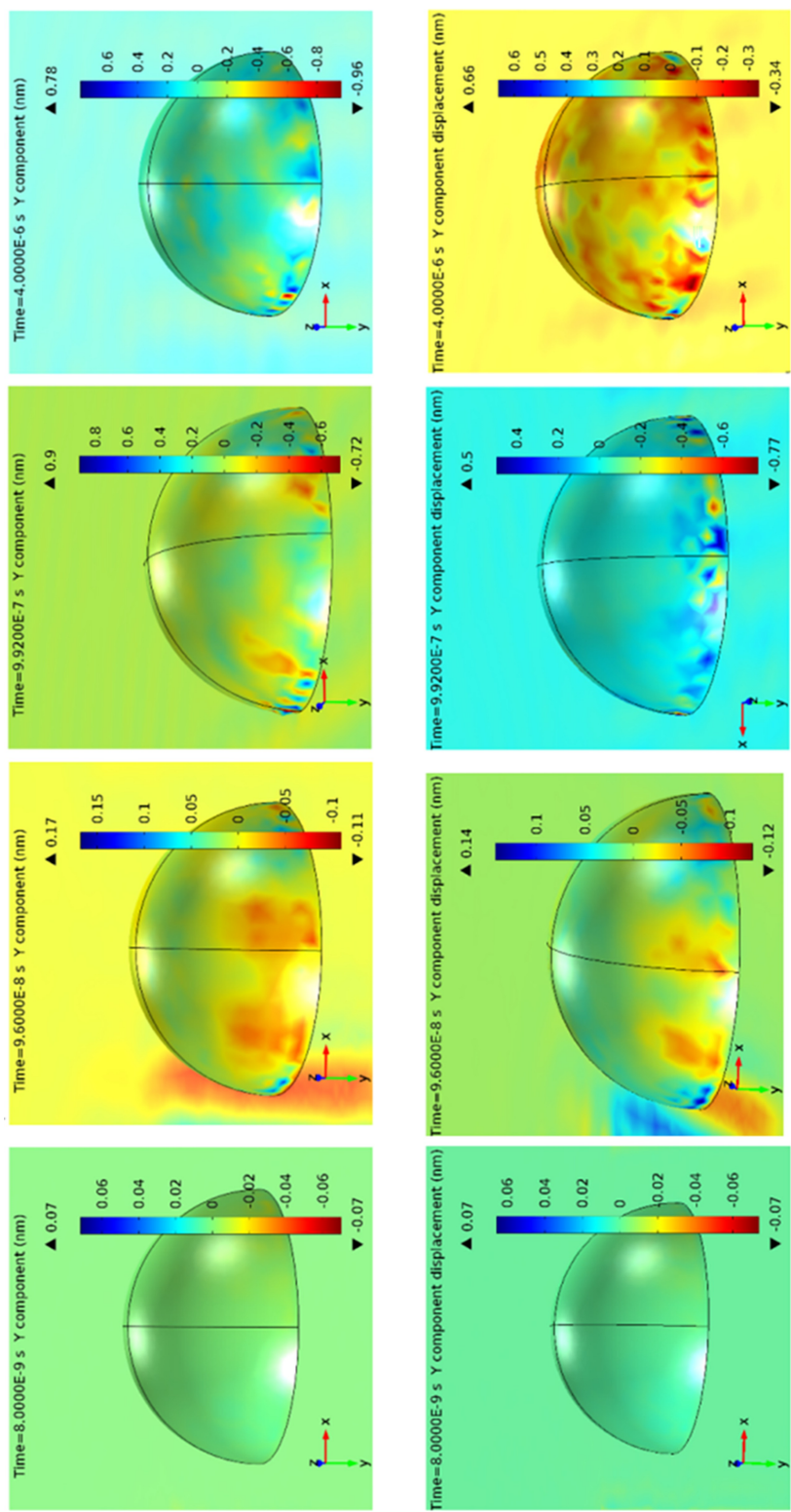


Figure 14.  
Mixing effect inside the droplet at different times.

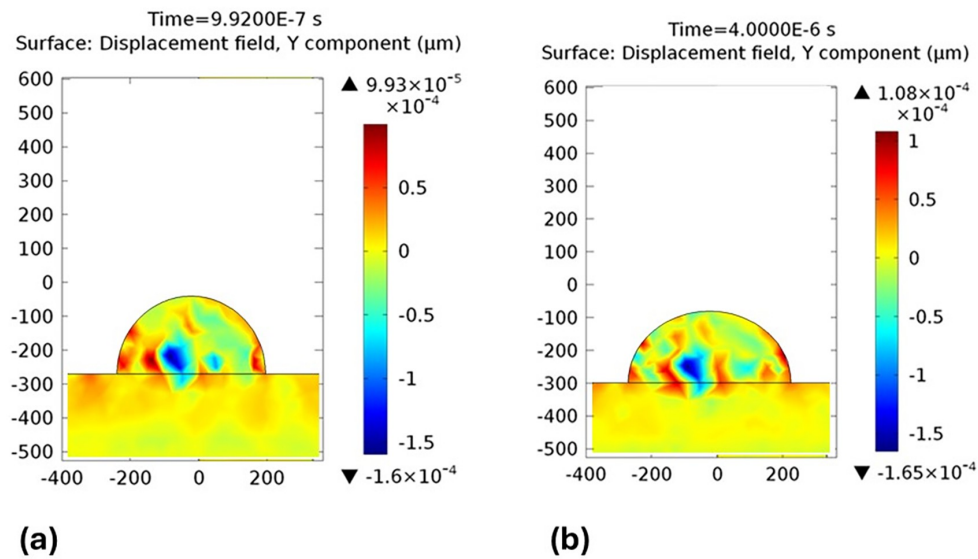


**Figure 15.**  
*Shows the results of the y-direction displacement inside the droplet due to SAW.*

## 5. Comparative discussion

The finite element-based COMSOL simulations reveal distinct advantages and are preferred among micromixer designs. The SAR designs are relatively simple compared to other designs and can produce effective results at low Reynolds numbers. The fractal micromixers provide higher mixing efficiency, but they are difficult to fabricate due to the complex nature of the geometry of the design.

The obstacle-based micromixers have the highest performance but cannot be practically fabricated due to their small dimensions and sharp edges. The active micromixer, with parametric optimization using Taguchi methods, offers a 99% mixing index that can be utilized for specific applications. **Table 1** represents the summary of all micromixers simulated, analyzed, and tested, some of them practically. The table shows the highest mixing for active micromixers and the lowest for SAR micromixers.

## 6. Applications of FEA in micromixing

Micromixers, designed and investigated through finite element simulations using COMSOL, have found applications in:

All these topics, from the application point of view, can be further explained in detail below:

### 6.1 Biomedical diagnostics like polymerase chain reaction (PCR)

FEA is an emerging, powerful tool to model fluid dynamics and thermal-related profiles in biomedical diagnostics, such as polymerase chain reaction (PCR)

| Serial no.      | Name of proposed micromixer | Type of micromixer    | Length of micromixer   | Width of micromixer                                     | Depth of micromixer                                                                |
|-----------------|-----------------------------|-----------------------|------------------------|---------------------------------------------------------|------------------------------------------------------------------------------------|
| 1               | YHRCSAR                     | SAR                   | 30 mm                  | 300 μm                                                  | 200 μm                                                                             |
| 2               | YESSAR                      | SAR                   | 30 mm                  | 300 μm                                                  | 200 μm                                                                             |
| 3               | YHCSSAR                     | SAR                   | 30 mm                  | 300 μm                                                  | 200 μm                                                                             |
| 4               | TVRMM                       | Obstacles based       | 8 mm                   | 200 μm                                                  | 200 μm                                                                             |
| 5               | Pure active micromixer      | Active                | 80 μm                  | Width of channel = 10 μm<br>Diameter of chamber = 30 μm | –                                                                                  |
| Reynolds number | YHRCSAR (Mixing index)      | YESSAR (Mixing index) | YHCSSAR (Mixing index) | TVRMM (Mixing index)                                    | Pure active micromixer (AMNO) [10 Hz, 0.5 V, 0.05 m/s] (optimal set of parameters) |
| 0.5             | 0.1203                      | 0.1102                | 0.0936                 | 0.982                                                   | Time (sec)    Mixing index                                                         |
| 1               | 0.1418                      | 0.1321                | 0.0821                 | 0.9178                                                  | 0    0.7839                                                                        |
| 10              | 0.2573                      | 0.2365                | 0.1837                 | 0.8044                                                  | 0.1    0.9891                                                                      |
| 20              | 0.3569                      | 0.3311                | 0.2295                 | 0.832                                                   | 0.2    0.9960                                                                      |
| 40              | 0.8065                      | 0.7751                | 0.7278                 | 0.9106                                                  | 0.3    0.9909                                                                      |
| 60              | 0.9591                      | 0.901                 | 0.8532                 | 0.9695                                                  | 0.4    0.9857                                                                      |
| 80              | 0.976                       | 0.9531                | 0.9021                 | 0.9588                                                  | 0.5    0.9787                                                                      |
| 100             | 0.9818                      | 0.9926                | 0.9436                 | 0.9705                                                  |                                                                                    |

**Table 1.**  
*Summary of micromixer results of the mixing index for a range of 0.1–100 Reynolds numbers.*

systems. The uniform mixing of reagents is critical for reaction efficiency, sensitivity, and reproducibility in PCR-based microfluidic devices. Simulation-based designs using FEA help in predicting temperature gradients, optimizing chamber design, and bridging the gap between theoretical reaction kinematics/kinetics and real-time PCR performance. Recent advances in droplet-based microfluidics using two-phase flow and digital control show that efficient mixing improves the amplification of nucleic acids and detection times [15]. Moreover, FEA allows the coupling of thermal cycling with convective flow to accurately replicate experimental PCR conditions. It is evident that the integration of FEA-based predictive numerical models with experimental validation accelerates the selection and utilization of better capability designs for POC PCR devices, thereby reducing prototyping, material, and labor costs. Some studies also highlight the role of FEA in understanding the concept of nanoparticle-assisted PCR and its mixing effect behavior for improved sensitivity [16]. Overall, in conclusion, we can say that FEA bridges theoretical fluid and heat transport with practical, real-time diagnostic performance.

## **6.2 Biodiesel production and esterification through efficient microreactors**

The efficient biodiesel synthesis in microreactors requires precise control of mixing and kinetic reactions. Thus, FEA is utilized by simulating coupled transport phenomena to successfully carry out the reaction. The esterification (reaction of ethanol with any oil, such as sunflower oil) and transesterification reactions depend mainly on the interfacial mixing of oil and alcohol, which is why micromixer design is very important. FEA-based numerical simulations have been used to optimize passive mixers like zig-zag, T-mixers, herringbone, obstacles, and M-shaped geometries, leading to a significant increase in biodiesel yield and production [17]. The numerical models help evaluate the pressure drop and mixing index for biodiesel preparation. Recent work by Shah et al. highlighted that computational methods, such as FEA, can be implemented for testing the Y-shaped Cantor fractal-shaped channel with material properties of ethanol and sunflower oil. The results are promising, as expected, and it can be tested in real-time for biodiesel preparation in the future [18]. Similarly, optimizing the catalysts and continuous microchannel designs for waste oil feedstocks is presented and discussed in detail in the review article by Osman et al. [19]. Thus, FEA helps bridge the gap between ideal chemical kinetics and real reactor operation in the industry by considering low Reynolds number-based turbulence, two-phase flow, and temperature gradients. This results in cost-effective prototyping and fabrication and reduces experimental trial-and-error-based designs. Hence, simulation-based FEA models for biodiesel microreactor designs improve energy efficiency and sustainability.

## **6.3 Pharmaceutical synthesis, drug delivery, and chemical reactions**

In pharmaceutical manufacturing industries, FEA is used to support the predictive modeling of chemical synthesis, controlled drug delivery, and release mechanisms. Accurate mixing is essentially required for uniform drug formulation, nanoparticle material, and chemical synthesis reaction kinetics. The FEA-based computational fluid dynamics (CFD) enables researchers to optimize micromixer geometries for high-throughput synthesis and nanocarrier encapsulation [20]. In drug delivery systems, FEA helps to simulate the injection forces, velocity of flow, capsule disintegration, and microfluidics-based implant release rate. This is useful for bridging theory with real-time in vivo performance, close to in vitro [21]. Some recent advances in self-driving power labs integrate FEA with artificial intelligence (AI) for automated reactions, reducing development time, and aiding in optimization. The uniform mixing predicted by simulation analysis ensures therapeutic safety. Coupling chemical reactions through chemical engineering concepts with FEA mechanical-based methods shortens validation cycles for modern pharmaceutical research and development (R&D). This integration consistently minimizes risk and is beneficial for accelerating drug development and experimentation.

## **6.4 Point-of-care devices testing for uniform fluid mixing**

The POC diagnostic instruments rely on precise and uniform fluid flow mixing to achieve trustworthy test results in resource-limited settings. FEA numerical simulations enable the micromixer design efficiency, and paper-based fluidic systems ensure reagent homogeneity. Recent numerical simulation-based FEA

studies on droplet microfluidics demonstrate that optimizing junction angles and channel geometries of Y-shaped channels enhances the mixing index up to 90% at low Reynolds numbers below 100 [22]. In paper-based microfluidics, localized electro-thermal FEA numerical models are used to enhance the reagent transport process without any additional integration of pumps. This is called a passive mixing mechanism [23]. Such models bridge the gap between the theoretical concept of fluid mechanics for laminar flows and real diagnostic test accuracy. By predicting steady-state and time-dependent mixing, FEA consistently reduces experimental failures and improves reproducibility. As POC devices become more complicated, simulation-driven designs will be key to translating laboratory precision into field-ready performance for real-time applications. Shah et al. tested their designs for biomedical purposes to some extent and planned to test them practically as a pilot project; the designs were tested and approved for future real-time use [24].

## **7. Challenges and future directions**

While FEA analysis provides accurate predictions of micromixing capability for real-time fabrication and testing challenges remain in bridging simulations with experimental realities and validation. As there are fabrication limitations, high-level accuracy of dimensions is required, and non-ideal fluid behavior must be accounted for, as some properties may be ideal but may vary in real-time. Additionally, similar scaling effects must be considered. Future research must explore hybrid micromixers that combine different ideas, such as fractal and obstacle-based designs, with machine learning algorithms for perfect optimization to accelerate and more accurately refine the design processes.

## **8. Conclusion**

The FEA is a powerful tool for the optimization of micromixer design and performance evaluation, especially for obtaining predictive results before practical fabrication. By combining numerical simulations with experimental validation, researchers can easily identify the optimal micromixer geometries that can be utilized for biomedical, chemical, synthesis, and thermal applications. Further advances in CFD techniques, along with optimization tools like Minitab software, further enhance the role of FEA in micromixing R&D. The cases studied and previous research have verified that mixing is an important process for biomedical and POC devices to produce uniform mixing for any chemical reaction, such as drug delivery, nanoparticles, etc. The concept is important as lab-on-chip devices have three processes: mixing, separation, and sorting. These three processes are either used individually or in combination in many devices, especially mixing, which is very critical and required for any kind of real-time POC device. The cases discussed show that FEA, using a numerical approach for predictive results, bridges the gap between theoretical and real-time applications.

## **Acknowledgments**

This work is supported by Key R&D Projects of Jilin Province of China (Grant Nos 20220201035GX, YDZJ202501ZYTS304 and 20230401099YY), Science Center for Gas Turbine Project (Grant No P2022-A-IV-002- 003), Key R&D Projects of the Ministry of Science and Technology of China (Grant Nos. 2017YFA0701200 and 2018YFB1107600) and National Natural Science Foundation of China (Grant No 51775237).

## **Author details**

Imran Shah<sup>1</sup>, Babar Hayat<sup>2\*</sup> and Shijun Ji<sup>2</sup>

1 Department of Mechatronics Engineering, Air University, Islamabad, Pakistan

2 School of Mechanical and Aerospace Engineering, Jilin University, Changchun, China

\*Address all correspondence to: babarhayat91@gmail.com

## **IntechOpen**

---

© 2025 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Balasuriya S. Dynamical systems techniques for enhancing microfluidic mixing. *Journal of Micromechanics and Microengineering*. 2015;259:094005
- [2] Ghahfarokhi NJ, Bayareh M, Nadooshan AA, Azadi S. Mixing enhancement in electroosmotic micromixers. *Journal of Thermal Engineering*. 2021;7(2):47–57
- [3] Liu YZ, Kim BJ, Sung HJ. Two-fluid mixing in a microchannel. *International Journal of Heat and Fluid Flow*. 2004;25(6):986–995
- [4] Hayat, B., et al., Investigation of heat transfer performance of double-walled pipes of heat exchanger for power plant applications. *International Communications in Heat and Mass Transfer*, 2023. **149**: p. 107115
- [5] Hayat, B., et al., Effect of surface roughness and temperature on thermal properties of double-walled pipes of heat exchangers. *International Communications in Heat and Mass Transfer*, 2025. **168**: p. 109413
- [6] Ward K, Fan ZH. Mixing in microfluidic devices and enhancement methods. *Journal of Micromechanics and Microengineering*. 2015;25(9):094001
- [7] Shamloo A, Mirzакhanloo M, Dabirzadeh MR. Numerical Simulation for efficient mixing of Newtonian and non-Newtonian fluids in an electro-osmotic micro-mixer. *Chemical Engineering and Processing - Process Intensification*. 2016;107:11–20
- [8] Park MG, Jang H, Lee SH, Justin Lee C. Flow shear stress enhances the proliferative potential of cultured radial glial cells possibly via an activation of mechanosensitive calcium channel. *Experimental Neurobiology*. 2017;26(2):71–81
- [9] Shah I, Kim SW, Kim K, Doh YH, Choi KH. Experimental and numerical analysis of Y-shaped Split and recombination micro-mixer with different mixing units. *Chemical Engineering Journal*. 2018;358:691–706
- [10] Shah I, Aziz S, Soomro AM, Kim K, SW K, Choi KH. Numerical and experimental investigation of Y-shaped micromixers with mixing units based on cantor fractal structure for biodiesel applications. *Microsystem Technologies*. 2020. DOI: 10.1007/s00542-020-05036-9
- [11] Kim K, Shah I, Ali M, Aziz S, Khalid MAU, Kim YS, et al. Experimental and numerical analysis of three Y-shaped split and recombination micromixers based on cantor fractal structures. *Microsystem Technologies*. 2019;(10):2953–2965
- [12] Irfan M, Shah I, Niazi UM, Ali M, Ali S, Jalalah MS, et al. Numerical analysis of non-aligned inputs M-type micromixers with different shaped obstacles for biomedical applications. *Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering*. 2021;09544089211051630
- [13] Shah I, Su Jeon H, Ali M, Yang DH, Choi KH. Optimal parametric mixing analysis of active and passive micromixers using Taguchi method. *Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering*. 2019;233(6):1292–1303

- [14] Shah I, Uddin E, Mubashar A, Sajid M, Younis MY, Samad H, et al. Numerical investigation of surface acoustic wave (SAW) interacting with a droplet for point-of-care devices. *The International Journal of Acoustics and Vibration* 2019;**24**:632–637
- [15] Barzoki AK, Mohseni A, Bazyar MM, Mohammadi K. Investigating mixing efficiency in droplets: A comprehensive study of numerical modeling and experimental testing in 3D-printed microfluidic devices. 2024. Available from: <https://arxiv.org/pdf/2401.08354> [Accessed: 2025-September-9]
- [16] Yao Q, Yang X, Wang Y, Wang J, Huang S, Song J, et al. Development and preliminary evaluation of a nanoparticle-assisted PCR assay for the detection of *Cryptosporidium parvum* in calves. *Animals* 2022;**12**:1953
- [17] Subramaniam K, Ky W, KH W, Chong CT, Ng JH. Enhancing biodiesel production: A review of microchannel reactor technologies. *Energies*. 2024;**17** (7):1652
- [18] Shah I, Aziz S, Soomro AM, Kim K, SW K, Choi KH. Numerical and experimental investigation of Y-shaped micromixers with mixing units based on cantor fractal structure for biodiesel applications. *Microsystem Technologies*. 2021;**27**(5):2203–2216. DOI:10.1007/s00542-020-05036-9
- [19] Osman AI, Nasr M, Farghali M, Rashwan AK, Abdelkader A, Al-Muhtaseb AH, et al. Optimizing biodiesel production from waste with computational chemistry, machine learning and policy insights: A review. *Environmental Chemistry Letters*. 2024;**22**(3):1005–1071. DOI: 10.1007/s10311-024-01700-y
- [20] Tom G, Schmid SP, Baird SG, Cao Y, Darvish K, Hao H, et al. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*. 2024;**124**(16):9633–9732. DOI: 10.1021/acs.chemrev.4c00055
- [21] Finite Element Analysis for Medical Devices & Biomaterials. Exponent [Internet]. [cited 2025 Sep 10]. Available from: [https://www.exponent.com/capabilities/finite-element-analysis-medical-devices-biomaterials?utm\\_source=chatgpt.com](https://www.exponent.com/capabilities/finite-element-analysis-medical-devices-biomaterials?utm_source=chatgpt.com)
- [22] Hajihadi Naghash T, Haghgoo AM, Bijarchi MA, Ghassemi M, Shafii MB. Performance of microball micromixers using a programmable magnetic system by applying novel movement patterns. *Sensors and Actuators, B: Chemical*. 2024;**406**:135403
- [23] Kunti G, Das SS, Pedireddi VM, Bhattacharya A, Chakraborty S. Energy-efficient micromixing in paper based devices mediated by the interplay of electrical and thermal fields. 2019 [cited 2025 Sep 10. Available from: <https://arxiv.org/pdf/1905.06972>
- [24] Irfan M, Shah I, Niazi UM, Ali M, Ali S, Jalalah MS, et al. Numerical analysis of non-aligned inputs M-type micromixers with different shaped obstacles for biomedical applications. *Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering*. 2022;**236**(3):870–880. DOI: 10.1177/09544089211051630

# Finite Element Method for Design of Flexible Tactile Sensors Toward Wearable Monitoring Applications

*Peng Wang, Yu Lv, Li Chen, Xiufeng Li, Xiaoyu Cao and Chuizhou Meng*

## Abstract

Flexible tactile sensors have recently emerged as a core component with potential for practical applications in wearable electronics, including health monitoring, motion recognition, human-machine interaction, and soft robotics. However, the traditional trial-and-error sensor development cycle of “material screening-device fabrication-performance testing” is constrained by the limitations of low efficiency, high cost, and deficient mechanistic insight. The finite element method (FEM), as a powerful numerical simulation technique, offers a viable approach to address these issues through discretization of complex geometries and establishment of coupled multiphysics equations encompassing mechanical, electrical, and thermal phenomena. This book chapter introduces the fundamental principles and reviews the state-of-the-art applications of FEM in the design of flexible tactile sensors, especially the ones with complex structuration. First, the general workflow of FEM modeling, selected strategies for mainstream software, and key techniques such as multiphysics coupling and constitutive behavior of materials are examined. Second, the role of FEM in the design optimization and mechanistic investigation of four typical tactile sensor categories, including the piezoresistive, capacitive, triboelectric, and temperature ones, is discussed with numerous representative progressing works. Third, the prospective advancements in FEM, including the development of more sophisticated multiphysics coupling models, integration of machine learning for the forward design of smart materials, expansion into system-level integration with reliability assessment, and promotion of sustainable design strategies, are outlined. In conclusion, the FEM method plays a pivotal role in advancing the next-generation flexible sensor technology toward high performance, integration, intelligence, and sustainability through mutual verification between virtual simulation and experimental validation.

**Keywords:** finite element method (FEM), flexible tactile sensors, multiphysics coupling, structural optimization, wearable electronics

## **1. Introduction**

The skin is known as the largest sensory organ of the human body, and can sensitively perceive external stimuli across diverse and wide ranges, including pressure (1–100 kPa), strain (0.1%–50%), and temperature (20–40°C). This naturally biological perception system offers valuable bioinspiration for the development of flexible tactile sensors [1–3], which have recently emerged as a core component with potential for practical applications in wearable electronics, including health monitoring, motion recognition, human–machine interaction, and soft robotics [4–6]. From the perspective of sensing targets, flexible tactile sensors can be broadly categorized into three types. Pressure sensors are primarily used for detecting physiological signals such as pulse and respiration, as well as motion status such as plantar pressure, and demand high sensitivity ( $>1$  /kPa) and low detection limits ( $<10$  Pa). The strain sensors can detect motions such as joint bending and muscle contraction, and require large strain ranges ( $>30\%$ ) and low hysteresis ( $<5\%$ ). The temperature sensors can convert the temperature-dependent electrical changes of materials into body or environmental temperature readings, and demand a high temperature coefficient ( $>1\%/^{\circ}\text{C}$ ) and strong anti-interference capability.

The challenge shared across all these sensor devices is achieving high performance and long-term stability while maintaining mechanical flexibility and biocompatibility. However, the traditional trial-and-error sensor development cycle of “material screening-device fabrication-performance testing” faces three major bottlenecks. The first one is low efficiency: optimization of device performance typically requires repeated adjustments of material ratios and structural parameters, and each experimental iteration can take several months. The second one is high cost: functional materials such as MXene and graphene, as well as precision fabrication equipment, are expensive; repeated trials lead to significant material waste. The third one is deficient mechanistic insight: experimental methods struggle to directly observe the coupled electro-mechanical behavior, such as conductive pathway evolution or ion migration under strain, restricting the theoretical guidance and structural optimization. To address these challenges, the FEM method can be used as a powerful and efficient numerical tool. FEM discretizes complex geometries into finite elements and establishes coupled multiphysics equations for mechanical, electrical, and thermal properties, enabling accurate prediction of material behavior and device responses. Its advantages in designing flexible tactile sensors are threefolded. The first one is accelerating the development: simulation allows pre-experimental performance prediction, for example, optimizing the active layer thickness or electrode spacing, thus reducing the experimental iterations by up to 60%. The second one is reducing the cost: virtual modeling can optimize material parameters without extensive physical prototypes, saving 30%–50% in material and equipment costs. The third one is gaining the mechanistic insight: FEM visualizes internal stress distribution, current density variation, and ion migration pathways within sensors, providing a theoretical basis for performance tuning [7–10]. As a result, FEM has been widely applied to the design of multiple types of flexible tactile sensors, such as piezoresistive, capacitive, triboelectric, and temperature sensors. When tightly integrated with experiments, FEM has significantly enhanced the controllability, performance, and research efficiency. In the field of wearable monitoring, FEM not only supports the structural optimization of sensing devices but also elucidates sensing mechanisms through multiphysics coupling analysis.

## 2. Fundamentals of FEM for flexible tactile sensors

The design of flexible tactile sensors involves complex interactions between mechanical deformation and electrical response, with device performance determined by the coupled relationship between material properties and structural configuration. As a critical tool for studying these material–structure–performance interactions, the FEM enables virtual simulation to provide pre-experimental structural optimization and performance prediction. This section introduces the simulation software selection, modeling workflow, and key technical considerations.

### 2.1 Comparison and selection of FEM software

The simulation of flexible tactile sensors often involves coupled multiphysics problems, including the mechanical, electrical, and thermal properties. Different FEM software platforms vary in accuracy, computational efficiency, and application scope. Currently, the most widely used platforms include ANSYS APDL, COMSOL Multiphysics, ABAQUS, and LS-DYNA [11–13]. Their core features are summarized in **Table 1**.

When selecting a simulation platform, the choice should align with the physical phenomena involved. If the research involves coupled thermo–electro–mechanical fields (e.g., the temperature–pressure dual-mode sensors), ANSYS APDL or COMSOL Multiphysics is preferred. For studies focusing primarily on mechanical performance or fatigue life prediction, ABAQUS is recommended. For dynamic response or transient signal analysis, LS-DYNA offers superior performance. Overall, ANSYS and COMSOL are the most widely adopted platforms in flexible tactile sensor research. Among them, COMSOL is favored for its user-friendly graphical user interface (GUI) and versatile multiphysics coupling, making it suitable for beginners and interdisciplinary collaborations. In contrast, ANSYS excels in handling complex models requiring high computational accuracy and convergence control.

| Software            | Advantages                                                                                                                     | Limitations                                                     | Applications                                        |
|---------------------|--------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|-----------------------------------------------------|
| ANSYS APDL          | High accuracy; supports nonlinear large-deformation and creep analysis; rich material library and custom constitutive modeling | High complexity and computational cost                          | Multiphysics coupling (pressure/resistive sensors)  |
| COMSOL Multiphysics | Excellent multiphysics coupling; user-friendly GUI; flexible meshing                                                           | Slower for highly complex 3D models                             | Porous structures, ionic migration, thermal sensors |
| ABAQUS              | Strong nonlinear mechanics and fatigue analysis; custom constitutive options                                                   | Weak in electrical simulation; often coupled with other solvers | Structural lifetime analysis of strain sensors      |
| LS-DYNA             | Efficient transient/dynamic response simulation                                                                                | Limited for static analysis                                     | Impact, vibration, and transient loading scenarios  |

**Table 1.**  
*Common FEM software for modeling and simulation of flexible sensors.*

## 2.2 General workflow of FEM modeling and simulation

The FEM simulation of flexible tactile sensors generally follows a standardized process: requirement definition → model construction → solution and verification → parameter optimization, ensuring both reproducibility and consistency with experimental results [13, 14]. (1) Requirement definition and parameter setting: after determining the sensor type (strain, pressure, or temperature), application scenario, and performance objectives, key parameters should be collected, including: (a) material properties: for example, polydimethylsiloxane (PDMS) elastic modulus 1.0–1.5 MPa, MXene conductivity  $10^4$ – $10^5$  S/m; (b) geometric parameters: for example, active layer thickness 10–100  $\mu\text{m}$ , electrode spacing 1–5 mm; (c) loading and boundary conditions: strain 0%–50%, pressure 0–100 kPa, voltage 0.5–1 V, temperature 25–37°C. (2) Geometry modeling and meshing: sensor geometries can be constructed directly in CAD software (e.g., SolidWorks) or reconstructed from image data. For composite or porous structures, material domains should be clearly partitioned and assigned appropriate properties. Critical regions (e.g., electrode interfaces) require refined meshes of 1–5  $\mu\text{m}$ , while the substrate can be coarsened to 10–20  $\mu\text{m}$  to balance computational efficiency and accuracy. (3) Boundary conditions and solver setup: in mechanical simulations, fixed and loaded boundaries are typically defined to model bending, compression, or stretching behaviors. In electrical simulations, voltage or current stimuli are applied across electrodes. The solution type may include steady-state analysis (for performance curves) or transient analysis (for response dynamics). A common convergence criterion is a force residual error below  $10^{-5}$ . (4) Post-processing and iterative verification: after solving, results such as stress distribution, current density, and impedance variation are extracted and compared with experimental measurements. If the deviation exceeds 10%, material parameters or mesh resolution should be adjusted for re-iteration. Additionally, parametric scans, such as varying the conductive filler content or layer thickness, can be conducted to identify optimal design configurations.

## 2.3 Multiphysics coupling and constitutive behaviors in FEM

The working mechanism of flexible tactile sensors essentially involves a force–electric multiphysics coupling process. Different sensing mechanisms correspond to distinct field interactions and mathematical descriptions. Resistive sensors: the external load causes deformation and rupture of the conductive network, resulting in a macroscopic change in electrical conductivity, which can be approximately expressed as  $\sigma = \sigma_0(1 + k\varepsilon)$ , where  $\sigma$  is the electrical conductivity under strain,  $\sigma_0$  is the initial conductivity,  $k$  is the gauge factor, and  $\varepsilon$  is the strain. Capacitive sensors: the applied pressure alters the electrode spacing or contact area, thereby changing the capacitance according to  $C = \varepsilon_0 \varepsilon_r \frac{A}{d}$ , where  $\varepsilon_0$  is the vacuum permittivity,  $\varepsilon_r$  is the relative dielectric constant,  $A$  is the effective electrode area, and  $d$  is the electrode distance. Temperature or ion-based sensors involve more complex thermo–electro–mechanical or ion–electro–mechanical coupling effects. In such systems, temperature variation or ion migration induces correlated changes in local potential distribution and mechanical deformation, which can be described by the Poisson–Boltzmann equation to capture the relationship between ion concentration and electrostatic potential [15].

In addition, flexible elastic materials typically exhibit nonlinear, viscoelastic, and even strain-dependent conductive behaviors; thus, the appropriate choice of constitutive model is essential. For elastomers such as PDMS and thermoplastic polyurethanes (TPU), Mooney–Rivlin or Neo–Hookean models are commonly employed to describe their nonlinear stress–strain relationships. For hydrogels or ionogels, due to their water content and viscoelastic polymer chain behavior, the Burgers model or generalized viscoelastic models are more suitable. For nanocomposite conductive systems, the Halpin–Tsai model can be used to predict the effective modulus, while the percolation theory describes the evolution of conductive pathways under varying strain conditions. By establishing a proper constitutive model, the microstructural parameters (such as filler distribution and crosslinking density) can be quantitatively correlated with the macroscopic response properties (such as sensitivity, response time, and stability), enabling a multiscale performance prediction from microscopic mechanisms to macroscopic behaviors. Therefore, the accuracy of constitutive modeling and multiphysics coupling settings is the core prerequisite for ensuring the physical reliability and experimental consistency of FEM simulations.

### **3. Advancements in FEM design of flexible tactile sensors**

#### **3.1 FEM design of resistive tactile sensors**

Resistive flexible tactile sensors have been extensively studied and applied in wearable monitoring, electronic skin, and human–machine interaction systems owing to their simple structure, easily readable signals, and rapid response characteristics. Their fundamental working mechanism is based on the change in the conductive network or contact resistance of the active layer under external mechanical stimuli, resulting in a measurable resistance variation. However, key performance indicators such as sensitivity, linearity, stability, and mechanical reliability are strongly dependent on the microstructural design of the active layer and the composite material properties. To address these dependencies, FEM provides an effective means of linking the structural mechanics with the electrical transport behavior. By virtually reconstructing the deformation and conduction processes, FEM enables the prediction of sensor response characteristics under various load conditions, offering valuable insights for structural optimization and performance enhancement.

##### *3.1.1 FEM design of resistive strain sensors*

By establishing the coupled mechanical–electrical models, FEM can visualize and quantify the stress distribution, current density evolution, and structural deformation during the stretching process, thereby providing theoretical support for optimizing material parameters, microstructural geometries, and device architecture.

##### *3.1.1.1 Bioinspired and microstructural design*

Bioinspired microstructural design has proven to be an effective strategy for enhancing both the sensitivity and structural stability of resistive strain sensors.

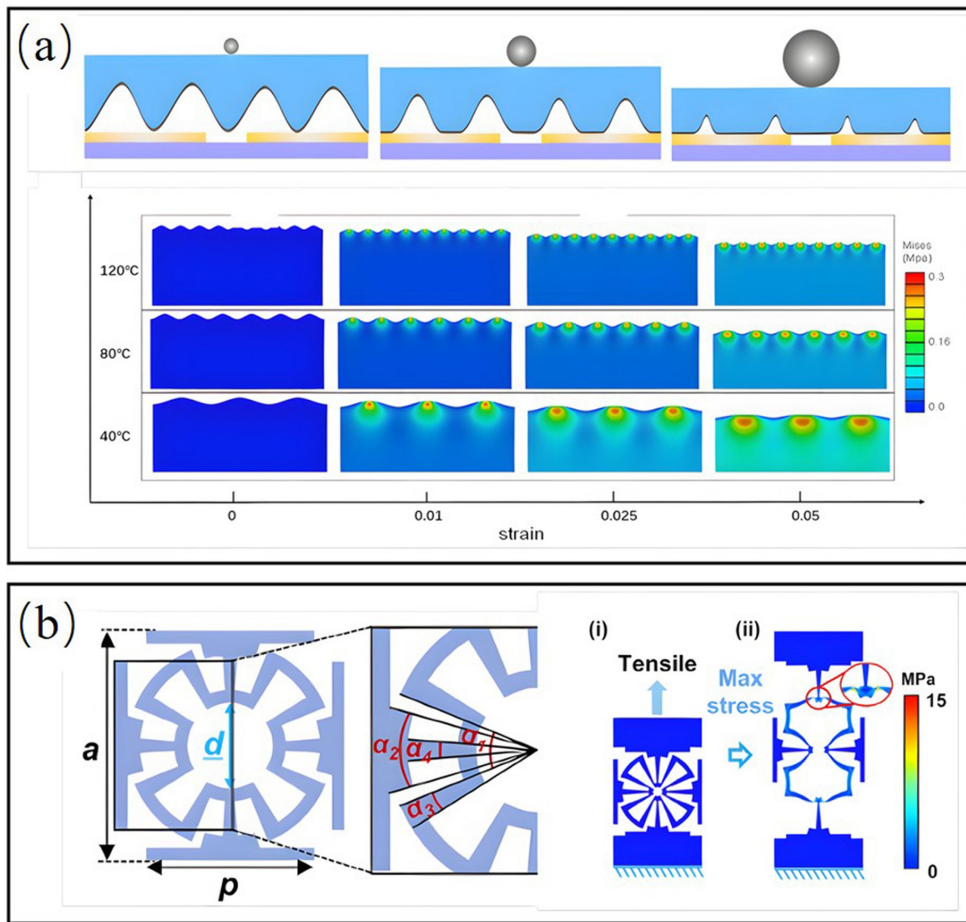


**Figure 1.** FEM on the bioinspired and microstructural design of resistive strain sensors. (a) Working principle diagram of the pressure sensors, and stress distribution of wrinkled GNWs annealed at different temperatures (40, 80, and 120 °C) under different strains ( $\epsilon = 0, 0.01, 0.025$ , and  $0.05$ ) (used with permission [16]). (b) Schematic representation of the circular auxetic design with detailed geometrical parameters and stress distribution of the structure at (i) 0% and (ii) 100% tensile strain (used with permission [18]).

Various natural systems – such as plant tracheids, skin wrinkles, and fish-scale textures – exhibit remarkable deformability and stress-dispersion capabilities, inspiring the construction of stretchable resistive architectures with improved mechanical and electrical performance.

For example, Tang et al. [16] developed a flexible pressure sensor based on a wrinkled graphene nanowall (GNW)/PDMS composite structure, where the wavelength and amplitude of the wrinkles were precisely tuned by controlling the annealing temperature. This approach enabled the fabrication of highly controllable and sensitive surface microstructures. Using ABAQUS, the team systematically simulated the stress distribution and contact area variations under different wrinkle geometries (see **Figure 1a**). Simulation results revealed that when the wrinkle wavelength exceeded  $20\ \mu\text{m}$  and the amplitude was greater than  $1.5\ \mu\text{m}$ , the variation

of conductive pathways during stretching was significantly enhanced, leading to an approximately 50% increase in sensitivity compared with flat structures. Furthermore, the FEM simulation uncovered reversible stress-transfer pathways within the wrinkled structure during cyclic deformation, elucidating the mechanism behind the sensor's low hysteresis and high repeatability under repeated loading. The simulation trends were consistent with experimental results, confirming the model's accuracy. The sensor achieved a high sensitivity of 59 /kPa in the 0–2 kPa range, maintained 4.8 /kPa in the 2–20 kPa range, exhibited a response time below 6.9 ms, and detected pressures as low as 0.2 Pa. Owing to its excellent dynamic performance, the device demonstrated strong potential for applications in plantar pressure monitoring, tactile imaging, and Braille recognition, providing an engineeringly feasible solution for electronic skin and prosthetic tactile systems.

In another work, Wei et al. [17] drew inspiration from the hierarchical porous architecture of plant xylem tracheids, constructing a carbon-based aerogel sensing layer composed of cellulose nanofibers and  $\text{Ti}_3\text{C}_2\text{T}_x$  (MXene), with chitosan (CS) incorporated as a bioinspired bridging phase to improve the flexibility and mechanical stability of the framework. The resulting three-dimensional (3D) porous conductive network was fabricated via freeze-drying and carbonization, achieving excellent compressibility and electrical conductivity. Using a 3D porous network FEM model, the team simulated the compressive deformation behavior under varying CS contents of below 10 wt%. The simulation results showed that when the CS content was 5 wt%, the area of stress concentration zones decreased by approximately 68%, while the structural recovery rate increased to about 92%, indicating that CS effectively buffered and redistributed local stress. The simulation trend closely matched experimental findings, verifying the predictive accuracy and design-guiding capability of FEM in optimizing composite ratios and mechanical response. The final sensor exhibited superior compressibility (stable over 10,000 cycles at 50% strain), high sensitivity (177.08 /kPa), and a fast response time (6.3 ms). These characteristics enabled high-performance pulse detection, respiratory monitoring, and joint motion recognition, highlighting its potential in intelligent wearables and biomedical monitoring.

In addition, Cho et al. [18] proposed a composite flexible strain sensor combining microcrack structures with auxetic metamaterials (negative Poisson's ratio) to overcome the limitations of conventional resistive sensors, which typically suffer from reduced sensitivity and nonlinear response under large strains. Using the COMSOL Multiphysics, the authors simulated various combinations of crack spacing ( $\sim 5\text{ }\mu\text{m}$ ) and prestrain ratios ( $\sim 30\%$ ) to elucidate the coupling relationship between geometric configuration, mechanical response, and resistance variation (see **Figure 1b**). The results revealed that with a  $5\text{ }\mu\text{m}$  crack spacing and 30% prestrain, the local stress concentration factor increased by approximately 2.3 times, leading to more pronounced disruption and reconstruction of conductive pathways, thereby amplifying the resistance change rate. Meanwhile, the auxetic structure's lateral expansion under tensile loading further enhanced strain sensitivity and linearity. The simulation results showed excellent agreement with experimental data, confirming that the coupling design of microcrack geometry and structural architecture is critical for performance optimization. The resulting strain sensor achieved ultrahigh sensitivity ( $\text{GF} > 1,200$ ), a broad strain range (0%–100%), and excellent stability, demonstrating wide applicability in motion recognition and soft-robotic strain feedback systems.

### 3.1.1.2 Process and interface-driven design

Beyond the structural morphology, the fabrication process and interfacial integrity play a decisive role in determining the long-term performance of flexible sensors. In particular, for inorganic or multilayer composite systems, stress concentration and material mismatch at microscopic interfaces often result in sensitivity degradation and signal drift.

For example, Han et al. [19] addressed the issue of sensitivity reduction and limited lifetime in conventional silicon-based strain sensors under large deformation by proposing a novel “inclined transfer crystal orientation modulation” strategy. In this design, single-crystal silicon nanowires (SiNWs) were employed as the sensing core, and their crystal orientation was programmably tuned through precise etching and transfer processes. The team established a 3D FEM model in Abaqus to simulate stress distribution and displacement response under different crystal orientations (e.g.,  $\langle 100 \rangle$  and  $\langle 111 \rangle$ ; see **Figure 2a**). The results revealed that the  $\langle 111 \rangle$  orientation exhibited the most uniform stress distribution, with the peak strain concentration reduced by approximately 35%, thereby mitigating conductive path fracture. The simulations further uncovered the relationship between crystal orientation and load-transfer pathways during cyclic tension–compression,

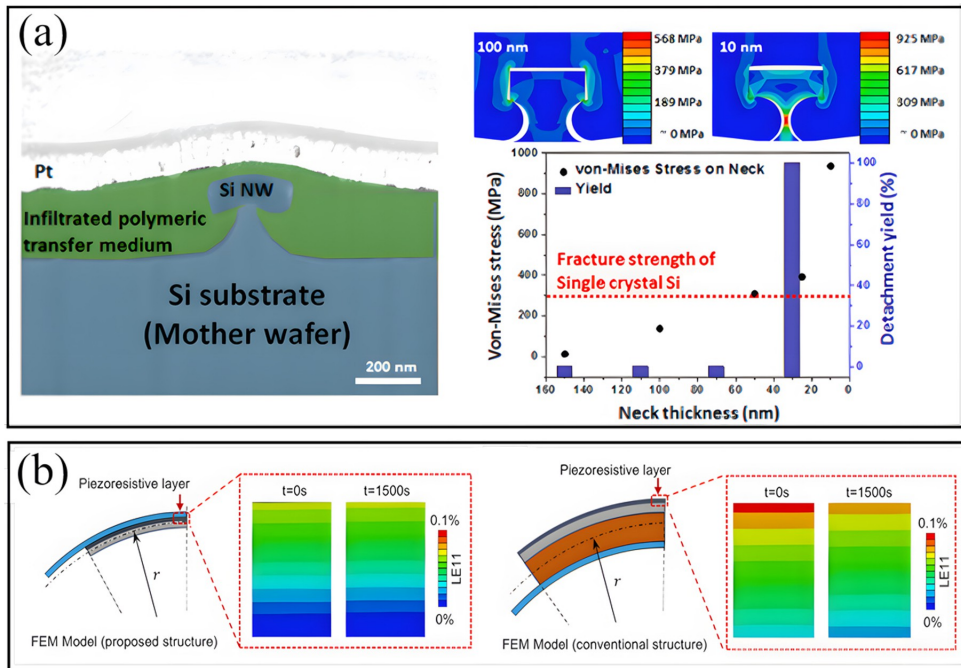


**Figure 2.** FEM on process and interface driven design of resistive strain sensors. (a) Cross-section SEM image of infiltrated polymeric transfer medium and grasping the 200 nm-wide silicon nanowires, top: simulation result of cross-sectional stress distribution in the SiNW at 12 nm displacement with neck thickness of 100 nm (left) and 10 nm (right), respectively, and bottom: change of von Mises stress on neck and experimentally measured detachment yield depending on neck thickness (dashed line: fracture strength of single-crystal Si) (used with permission [19]). (b) Simulation results of the change in surface stresses of proposed bending sensor, and the change in surface stresses of conventional structured sensor using adhesive induced by stress relaxation over time (used with permission [20]).

providing a quantitative link between crystallographic direction and mechanical stability. Experimental data aligned well with the FEM predictions, confirming the superior stress homogenization and sensitivity retention of the  $\langle 111 \rangle$ -oriented structure. This design achieved a 42% increase in normalized resistance change under 1%–3% strain and nearly tripled the fatigue lifetime. The study clearly highlights the predictive and instructive power of FEM in semiconductor microstructure process optimization, offering an effective route for developing high-sensitivity and long-lifetime flexible electronic sensors.

In another work, Takamatsu et al. [20] approached the challenge from an interface engineering perspective and developed an adhesive-free ultrathin silicon strain sensor based on Au–Au cold welding technology. This strategy effectively addressed the common problem of signal drift and fatigue failure caused by elastic modulus mismatch in traditional adhesive-bonded flexible electronics. Using the Abaqus FEM simulations, the team compared the stress distribution and relaxation behavior of conventional adhesive-layer bonding structures and Au–Au cold-welded interfaces under cyclic tensile loading (see **Figure 2b**). The results indicated that the maximum stress peak at the Au–Au interface was reduced by  $\sim 60\%$ , while the interfacial slip displacement was nearly halved (from  $\pm 15\%$  to  $\pm 3\%$ ), thereby suppressing signal drift effectively. The simulation further revealed that, in the adhesive-bonded structures, the large modulus differences lead to non-uniform energy dissipation during repeated stretching; in contrast, the adhesive-free cold-welded interface enables continuous stress transfer and load homogenization. Experimental findings were highly consistent with the FEM results, confirming the superior stress-transfer capability and long-term stability of this bonding strategy. The final device maintained high repeatability and stable output over 10,000 loading cycles, with an error margin within 0.8%, demonstrating significant potential for wearable health monitoring and precise motion detection applications.

#### *3.1.1.3 Fabric textile network design*

In wearable applications, strain sensors are often integrated into garments or textile substrates to monitor human motion, posture, and physiological signals. Such systems require not only high sensitivity and flexibility of individual sensing units but also overall structural integrity and signal consistency under large-scale bending and torsional deformation. FEM demonstrates particular advantages in these complex systems by enabling mechanical prediction and optimization of multi-layer structural coupling with the human body surface.

For example, Gupta et al. [21] proposed an “all-knitted and integrated” flexible strain-sensing system, in which the conductive fibers were embedded into an elastic textile substrate through knitting techniques to form a large-area two-dimensional strain detection network. This design maintained high stretchability while achieving intimate skin conformity and excellent signal stability. Using FEM, the research team established a “skin–fabric–sensing layer” composite model to analyze the stress distribution and signal response under varying bending radii (15–40 mm) and stretching directions. Simulation results revealed that when the sensing units were aligned along the knitting direction, stress concentration regions were minimized, and signal deviation remained below 5%. In contrast, the crosswise alignment increased local shear stress by approximately 18%, leading to nonlinear signal amplification. This study was the first to quantitatively correlate fabric alignment

patterns with strain-sensing performance, providing a theoretical foundation for structural optimization of wearable systems. In practical applications, the proposed system was implemented in a flexible data glove, capable of real-time detection of finger bending and gesture movements. The measured outputs exhibited over 90% consistency with FEM simulation results. Moreover, the device maintained stable performance under tensile strains exceeding 100%, demonstrating remarkable stretchability, high integration, and excellent adaptability for wearable applications. In practical validation, the system successfully achieved real-time monitoring of knee joint bending angles and gait patterns, with output signals exhibiting excellent agreement with simulation results. The FEM analysis not only identified the optimal arrangement of sensing units but also elucidated the energy transfer pathways and structural coupling behaviors of the textile substrate under mechanical loading. These insights established a fundamental framework for advancing wearable electronics design from the material level to the system level.

### *3.1.2 FEM design of piezoresistive pressure sensors*

Piezoresistive pressure sensors primarily transduce external pressure stimuli into electrical signals through the rearrangement of conductive networks, closure of porous structures, or variation of contact area within the sensing layer. Their key performance indicators, including sensitivity, linear operating range, and anti-interference capability, are strongly dependent on structural configuration, compressive modulus of the materials, and interfacial adhesion quality. FEM serves as a powerful tool in enabling quantitative visualization of stress distribution and current density evolution under compressive deformation. It can also predict the energy absorption behaviors of porous architectures and reveal the multi-scale mechanical coordination mechanisms in hierarchical structures. Through such simulations, FEM provides a direct pathway for translating microscopic structural design into macroscopic performance optimization.

#### *3.1.2.1 Porous structural design*

Porous architectures, known for their exceptional compressibility and energy dissipation characteristics, form the foundation for achieving high-sensitivity pressure sensing. The microstructural morphology and fabrication process of these porous materials directly influence the elastic recovery behavior and the evolution of conductive pathways within the sensing layer.

For example, Wang et al. [22] proposed a TPU/carbonized wood aerogel composite system to overcome the limitations of conventional wood-based sensors, which often exhibit single-functionality and poor triboelectric performance. Using balsa wood as a natural template, the team performed delignification and freeze-drying to obtain a highly porous wood aerogel, followed by carbonization at 800°C to produce a conductive scaffold with a porosity of 91.7%, while retaining the intrinsic “wave-like layered porous structure.” Subsequent impregnation with TPU solution yielded the elastic and conductive composite. To analyze its sensing mechanism, the researchers employed COMSOL Multiphysics to establish a vertical contact-separation triboelectric nanogenerator (TENG) model for the composite system, inputting material parameters such as the elastic modulus of TPU (0.38 MPa) and the dielectric constant of polytetrafluoroethylene (PTFE) (2.1). FEM

simulation revealed that the 3D conductive network effectively facilitates charge transport, and its rough porous surface increases the contact area, leading to simulated open-circuit voltage ( $V_{oc}$ ) and short-circuit current ( $I_{sc}$ ) values with less than 2% deviation from experimental data. In resistive sensing mode, FEM further demonstrated rapid pore closure and the fast formation of conductive paths under low pressures, corresponding to a sensitivity of 174.83 /kPa, again consistent with experimental observations. The multifunctional potential of this design was experimentally validated: a single-point sensor operated as a gradient dimming switch, while a  $3 \times 3$  array achieved 100% accurate password recognition. Moreover, the TENG enabled a self-powered gait and fall detection, with a power density of 5.64 W/m<sup>2</sup>. FEM modeling provided critical guidance throughout the design process, from material optimization and mechanical response range prediction to overall device performance forecasting.

In another work, Wang et al. [23] addressed the challenges of graphene aerogel collapse and low drying efficiency by proposing an innovative “sodium alginate (SA)–metal ion crosslinking + solvent-exchange dehydration” strategy for rapid gelation and structural stabilization. Using graphene oxide (GO) as the base, the researchers introduced a SA and EDTA–Ca<sup>2+</sup> hybrid crosslinking system to form foam hydrogels supported by air-bubble templates. Subsequent ethanol/ethyl acetate solvent exchange and thermal reduction at 1,000°C produced structurally stable r-SGS-Ca aerogels with 80% volume retention and tolerance for up to 80% compressive strain. A two-phase (hydrogel–solvent) mass transport FEM model was constructed in COMSOL to simulate mass migration and volumetric shrinkage during dehydration, driven by chemical potential gradients. The model revealed that rapid outer-layer dehydration and inner-bubble counter-support effectively mitigate internal stress accumulation. The simulated results showed less than 2% deviation from experimental data for both volume and mass change. Furthermore, the model identified an optimal SA concentration of 7.5 mg/mL, ensuring uniform stress distribution and mechanical stability. The FEM-optimized r-SGS-Ca aerogel-based piezoresistive sensor demonstrated a sensitivity of 4.89 /kPa over a 0–6 kPa range and exhibited excellent long-term cyclic stability. Its applicability was validated in wrist pulse monitoring, laryngeal swallowing detection, and motion-pad posture recognition, showcasing broad potential in wearable physiological sensing and intelligent motion detection.

#### *3.1.2.2 Hierarchical structural design*

Hierarchical and multiscale architectures can realize graded responses across different pressure regimes, thereby achieving both high sensitivity and a broad detection range.

For example, Wang et al. [24] proposed a “hierarchical porous sensing unit + multimodulus framework” concept, employing fully 3D printing to fabricate a stretchable piezoresistive sensor and using FEM to verify the rationality of the multilayer design. The material system consisted of three printable inks: a PDMS substrate layer (0.98 MPa) providing tensile support; a highly conductive and stretchable TPU/85% Ag electrode layer; and a TPU/carbon black (CB)/NaCl sensing layer forming a hierarchical macro–micro–nano porous network for multiscale strain regulation. A 3D FEM model incorporating all three layers was established, with corresponding mechanical parameters (electrode: 2.96 MPa; sensing layer: 0.38 MPa; Simulations under 50% in-plane strain and 10 Pa–800 kPa

out-of-plane pressure revealed that the multimodulus configuration effectively distributed mechanical stress i.e., about 90% of tensile stress was absorbed by the substrate, while the sensing layer experienced only 5% strain. The simulated resistance variation (<7%) matched experimental data. Under compressive loading, staged pore closure yielded a three-segment sensitivity response ( $R^2 > 0.98$ ), and FEM predicted six layers as the optimal configuration balancing sensitivity and stability. The resulting sensor could detect pressures from 10 Pa to 800 kPa, exhibiting exceptional durability (stable performance over 10,000 cycles). Application tests demonstrated precise pulse waveform monitoring on the wrist, rapid response in robotic fingertips outperforming commercial FSR400 sensors, and spatial pressure imaging in a 3D-printed  $4 \times 4$  array (e.g., chess piece recognition). This study highlights FEM's crucial role in structural validation and optimization of flexible sensor architectures.

In another work, Yan et al. [25] addressed the challenge of the narrow detection range in MXene-based pressure sensors by introducing a multilayer hierarchical composite design consisting of an inverted-cone PDMS surface layer, a MXene/multiwalled carbon nanotubes (MWCNTs)–PDMS conductive layer, and a  $\text{SiO}_2$  microsphere interfacial barrier layer. FEM was employed to optimize microsphere geometry and structural configuration, enabling both high sensitivity and wide detection range. The conductive layer, composed of  $\text{Ti}_3\text{C}_2\text{T}_x$  (MXene), MWCNTs, and PDMS, formed a continuous conductive network, while the dual-scale  $\text{SiO}_2$  spheres (20 nm and 30  $\mu\text{m}$ ) acted as multicontact support points. A three-layer FEM model was established with input moduli of 0.8 MPa (MXene/MWCNTs–PDMS) and 72 GPa ( $\text{SiO}_2$  layer), simulating electrical tunneling and contact behavior under 0–1,228.75 kPa pressure. The simulations revealed a three-stage response mechanism: at low pressure (0–1.25 kPa), microspheres maintained stable contact; in the mid-range (1.25–714.75 kPa), stress concentration at the inverted-cone interfaces induced a sharp contact area increase, yielding ultrahigh sensitivity (9,041.8 /kPa); and at high pressure (>714.75 kPa), the contact area saturated, maintaining a still-high sensitivity (1,596.2 /kPa). The simulated and experimental data showed a strong correlation ( $R^2 > 0.98$ ). FEM analysis further indicated that the dual-scale microspheres reduced peak stress in the conductive layer by  $\sim 42\%$  and maintained  $> 90\%$  stress retention after 5,000 cycles, significantly enhancing fatigue resistance. Application tests demonstrated accurate detection of finger bending ( $0^\circ$ – $120^\circ$ ), throat swallowing motions, and robotic recognition of materials with different stiffnesses, underscoring FEM's value in multiscale interface tuning and sensor performance design.

In addition, Zhang et al. [26] sought to overcome the issue of signal saturation and narrow linear range in porous flexible sensors by proposing a TPU/c-MWCNTs/PDMS gradient porous structure and employing FEM to elucidate a “macro–micro pore relay closure” mechanism that simultaneously improved sensitivity and linearity. The sensor comprised an upper macro-porous TPU layer (35–40  $\mu\text{m}$  pores) and a lower micro-porous layer (7–10  $\mu\text{m}$  pores) embedded with 0.1 wt% c-MWCNTs to form a conductive network. A 1 wt% PDMS coating enhanced mechanical robustness and humidity resistance. FEM simulations using material parameters (TPU: 2.8 MPa; c-MWCNTs/PDMS: 4.2 MPa) under 0–200 kPa compression revealed that large pores closed first in the low-pressure regime (0–60 kPa, sensitivity = 78.41 /kPa), while micro-pores engaged at higher pressures (60–200 kPa), extending linearity up to 200 kPa ( $R^2 > 0.998$ ; see **Figure 3**). PDMS modification improved compressive resilience by 56%, limiting resistance drift to

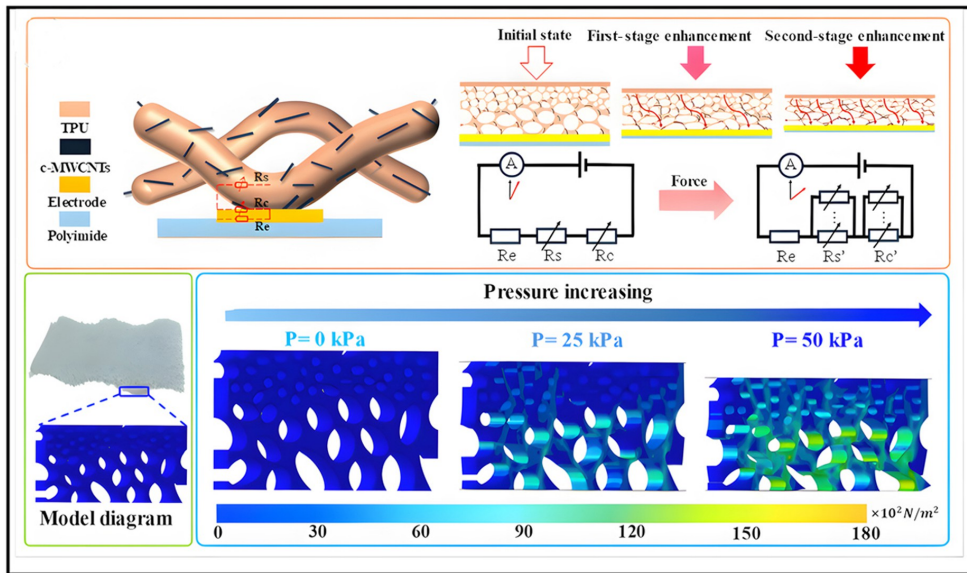


**Figure 3.**  
 FEM on the hierarchical structural design of flexible piezoresistive pressure sensors. Illustration of the sensing mechanism of the sensor, and simulation results showing the stress distribution of the sensor under different pressures (used with permission [26]).

< 6% after 10,000 cycles and suppressing humidity interference by  $\sim 65\%$ . The FEM-optimized  $4 \times 4$  tactile array, integrated with a convolutional neural network (CNN), enabled real-time Braille signal recognition (resolution = 8 Pa, character accuracy = 94.3%, voice transcription accuracy = 93.9%). Additionally, it effectively monitored wrist pulse and knee motion, demonstrating strong potential in assistive and wearable healthcare applications.

Furthermore, Gabbi et al. [27] addressed the issue of “unclear piezoresistor layout and limited sensitivity” in flexible paper-based piezoresistive sensors by proposing an FEM-based structural optimization approach to achieve low-cost and systematic performance enhancement. In this study, standard A4 printing paper was used as a flexible substrate, and graphite served as the piezoresistive material. A square diaphragm structure with four fixed edges was designed. Using the ANSYS AIM platform, the team constructed a 3D FEM model employing SOLID226 tetrahedral elements (mesh size: 1  $\mu\text{m}$ ) and applied uniform vertical pressure to analyze the stress distribution under different aspect ratios ( $a/b = 1.0$ – $2.0$ ). Simulation results revealed that when  $a/b = 1.0$ , the maximum stress occurred at the mid-edge region of the diaphragm, reaching  $2.729 \times 10^9$  Pa; when  $a/b = 2.0$ , the maximum stress decreased to  $2.151 \times 10^9$  Pa. Further analysis indicated that depositing the piezoresistor in the high-stress region increased the relative resistance change per unit pressure by approximately 30%. Experimental validation confirmed a pressure resolution of 1 Pa and a response time below 50 ms. The fabricated flexible paper-based sensor could be conformally attached to the wrist for pulse monitoring (distinguishing systolic and diastolic peaks), integrated into insoles to identify gait variations, and used as a flexible pressure switch for human–machine interaction. This work fully demonstrates the quantitative guidance provided by FEM in geometric parametric design and

piezoresistive structure optimization, offering a replicable and low-cost methodology for flexible electronic device design.

### **3.2 FEM design of flexible capacitive pressure sensors**

As an essential branch of flexible tactile sensors, capacitive pressure sensors possess inherent advantages such as high sensitivity, low power consumption, and simple structural configuration. These features have enabled their widespread application in wearable electronics, electronic skin, and human-machine interaction systems. The fundamental working mechanism relies on variations in capacitance induced by external pressure, which alter the electrode spacing or the dielectric constant within the sensing unit. In recent years, intensive research efforts have been directed toward improving the sensitivity, linearity, and stability of capacitive pressure sensors through structural optimization, electrode patterning, and material engineering. FEM has been extensively introduced as a powerful analytical tool for multiphysics simulations, enabling in-depth investigation of mechanical-electrical coupling mechanisms and providing quantitative guidance for device performance optimization. Among these, ionic supercapacitive pressure sensors have attracted particular attention, featuring two flexible electrode layers that sandwich a high-capacitance ionic dielectric. Upon external loading, the reduction in inter-electrode distance or the increase in effective contact area induces a pronounced change in capacitance, thereby enabling efficient transduction of mechanical stimuli into electrical signals. To further enhance sensitivity and conformability, researchers have incorporated advanced architectures such as MXene nanomaterials, conductive foams, and microarray-structured dielectrics, achieving notable improvements in response speed, detection range, and operational stability.

#### *3.2.1 Surface contact interface optimization*

Mishra et al. [28] conducted research on flexible capacitive pressure sensors. These sensors use Au, Ag, Cu, Al, silver nanowires (AgNWs), and other materials as electrodes, and air, vacuum, PDMS, Ecoflex, polyurethane (PU), graphene-based composites, porous elastomers, wrinkled materials, microstructured elastomers (micropillars, micropylamids, tilted micropillars), ionic liquids, and other substances as dielectrics. Fabricated via traditional micromachining techniques (such as surface/bulk micromachining, photolithography, and KOH etching), the sensors feature a unique parallel-plate capacitor structure. In the research, FEM software was employed to simulate diaphragms of different shapes (circular, square, rectangular, etc.), yielding behavioral laws including the sensor's mechanical sensitivity, electrical sensitivity, stress concentration distribution, resonant frequency, and response time, as well as capacitance-pressure curves and deformation-pressure curves (see **Figure 4a**). The study clarified the influence of parameters such as diaphragm shape, micropillar spacing/sidewall angle, and dielectric layer thickness on sensitivity, nonlinearity, and hysteresis characteristics, with the simulation results being mutually verified by experimental test data. Ultimately, the developed flexible capacitive pressure sensors underwent application validation in healthcare, robotics, consumer electronics, transportation, and marine

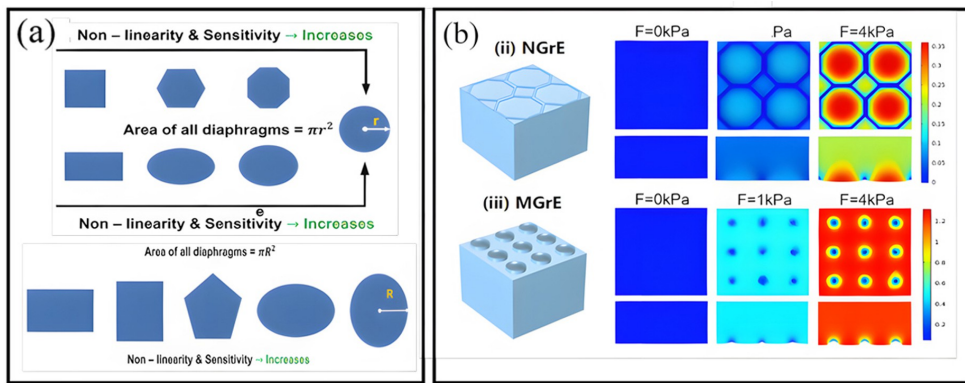


**Figure 4.** FEM on the surface contact interface optimization of capacitive pressures sensors. (a) Effect on the nonlinearity and sensitivity when the diaphragm shape changes from square to circle (with sides from 4, 6, 8, and  $\infty$ ) and rectangle to circle (through the elliptical shape) while keeping the overlapping area same (used with permission [28]). (b) Finite-element simulation showing the local stress distribution and the deformation of NGrE and MGrE (used with permission [29]).

fields, and hold significant application potential in aerospace, wearable devices, and other scenarios.

In another work, Yang et al. [29] utilized graphene as the core electrode material and fabricated smooth graphene electrodes (SGrEs), nanostructured graphene electrodes (NGrEs), and microstructured graphene electrodes (MGrEs) via polymethyl methacrylate (PMMA)-mediated transfer, ultraviolet-curable adhesive (UVA)-mediated transfer, and photolithography-wet etching micromachining, respectively. They then assembled a sandwich-structured flexible capacitive pressure sensor with PDMS as the dielectric layer. Among these components, the MGrEs, featuring controllable multiscale roughness ranging from the nanometer to micrometer scale, serve as the unique structural core of the sensor. The research employed COMSOL Multiphysics 5.3, a software for FEM, constructing a 3D mechanical–electrical coupling model for the interface system composed of NGrEs/ MGrEs and the PDMS dielectric layer (see **Figure 4b**). This model simulated the interface behavior under pressures ranging from 0 to 4 kPa, yielding the stress distribution in the electrode–dielectric contact region, the variation law of the displacement field, and the relative capacitance change rate ( $\Delta C/C_0$ ) curves. It also clarified the principle that MGrEs enable larger changes in electrode spacing and superior sensitivity due to their higher compressibility. The simulation results were mutually verified with experimental data: for instance, simulations indicated that the sensitivity of MGrE-based sensors was significantly higher than that of NGrE and SGrE-based counterparts, and experiments confirmed that the MGrE sensor achieved a sensitivity of 3.19 /kPa (319 times that of SGrE sensors), while the sensor with a symmetric double-MGrE structure reached an even higher sensitivity of 7.68 /kPa. This sensor has successfully undergone application validation in insect crawling detection (capturing subtle pressure changes during insect movement), wearable health monitoring (e.g., detecting the  $P_1$ ,  $P_2$ , and  $P_3$  peaks of radial artery pulse waves at the wrist), robotic tactile force feedback (integrated into smart gloves to perceive pressure when grasping water cups of different weights), and touch

localization using  $4 \times 4/3 \times 3$  sensor arrays. It holds significant application potential in wearable health devices, bionic electronic skin, human-machine interaction systems, and intelligent robotic tactile sensing scenarios.

Besides, Zhong et al. [30] developed a flexible capacitive pressure sensor using porous PDMS doped with high-concentration CNTs as the electrode material and an ultra-thin porous polyvinylidene fluoride (PVDF) fiber membrane as the dielectric layer. The porous electrodes were fabricated via the NaCl template method (by removing NaCl particles to form a porous structure), while the PVDF dielectric layer was prepared through electrospinning. The sensor's unique feature lies in its porous conductive electrode design, which enables continuous variation of the contact area between the electrode and the dielectric layer. For the study, ANSYS Explicit software combined with Unigraphics NX for modeling was employed to construct a mechanical simulation model of the porous electrode, simulating its compressive deformation behavior and the variation law of contact area under pressure. This yielded correlation curves between pressure and electrode deformation, as well as between pressure and contact area, and clarified the core principle that stress concentrates around the pores under low pressure, while the porous electrode gradually compacts and the contact area continuously increases with increasing pressure. The simulation results were mutually verified with experimental data, both confirming that the controllable deformation behavior of the porous electrode can significantly enhance sensor sensitivity. Experiments showed that the sensor achieved a sensitivity of 1.033 /kPa in the range of 0–1 kPa and maintained a high sensitivity of 0.23 /kPa even in the range of 10–30 kPa. This sensor has been successfully applied in human physiological signal monitoring (such as detecting the P-wave, T-wave, and D-wave of wrist pulses) and muscle activity monitoring (including finger bending, muscle contraction during dumbbell curls, and judging the degree of muscle activation during squats). Additionally, a  $3 \times 3$  sensor array was constructed to realize pressure distribution mapping of object shapes and weights. It holds significant application potential in wearable health monitoring devices, intelligent motion monitoring systems, human-machine interaction, and other scenarios.

### *3.2.2 Internal structural interface verification*

Zhang et al. [31] fabricated a sandwich-structured flexible capacitive pressure sensor using Ecoflex micropylamid arrays modified with CNTs as the electrode material and porous PDMS (prepared via the sucrose particle porogen method) as the dielectric layer. The micropylamid electrodes were shaped by template molding, while the porous dielectric layer was prepared through spin coating combined with ultrasonic cleaning. The sensor's unique feature lies in the tunable duty ratio (ranging from 1:2 to 1:9) and height (ranging from 60 to 90  $\mu\text{m}$ ) of the micropylamid array electrodes. FEM software was employed to construct mechanical simulation models for micropylamid array electrodes with different structural parameters, simulating their stress distribution and deformation process under pressure. This yielded the core principle that “the larger the duty ratio and height of the micropylamid array, the higher the sensor sensitivity,” clarified the deformation behavior wherein stress concentrates at the tips of the micropylamids under low pressure and the micropylamids gradually flatten under high pressure, and obtained stress-deformation curves corresponding to different pressures (0–10 kPa).

The simulation results were mutually verified with experimental data: for example, the simulation conclusion that micropylramids with a high duty ratio exhibit more significant deformation was consistent with the experimental result that the sensor with a duty ratio of 1:9 achieved a sensitivity of 2.24 /kPa. This sensor has been successfully applied in human physiological signal monitoring (such as detecting the P-wave, T-wave, and D-wave of wrist pulses) and motion state monitoring (including rope skipping counting, recording muscle contractions during dumbbell curls, and judging the standardization of squat movements). Additionally, a 4 × 4 sensor array was constructed to realize touch localization (e.g., identifying the contact positions of nails and nuts). It holds significant application potential in sports wearable devices, health monitoring instruments, human-machine interaction systems, and other scenarios.

In another work, Guo et al. [32] assembled a sandwich-structured flexible capacitive pressure sensor using PDMS/aluminum oxide ( $\text{Al}_2\text{O}_3$ ) composite as the dielectric material and conductive copper foil as the electrode material. The fabrication process involved preparing a bionic honeycomb-structured mold via 3D printing, followed by casting, curing, and surface conductive treatment. The unique core of the sensor lies in its dielectric layer with a high-porosity bionic honeycomb microstructure, which balances high sensitivity and a wide pressure sensing range. FEM software was utilized to construct a mechanical-electrical coupling simulation model for the bionic honeycomb composite dielectric layer, simulating the stress distribution, compressive deformation behavior, and capacitance variation law of the dielectric layer under different pressures. This yielded the correlation curve between pressure and relative capacitance change rate, clarified the regulatory mechanism of honeycomb porosity and pore size on sensor sensitivity and response speed, and revealed the deformation characteristics wherein stress concentrates at the joints of honeycomb cell walls under low pressure and the cell walls gradually compact under high pressure. The simulation results were mutually verified with experimental data: for instance, the simulation prediction that the high-porosity honeycomb structure exhibits superior sensitivity was consistent with the experimental results showing the sensor achieved a high sensitivity of 42.6 /kPa and ultra-fast response/recovery times of 21 and 22 ms, respectively. The sensor has successfully undergone fast response testing and repeatability verification for different pressure signals and can be applied in human physiological signal monitoring (such as pulse and respiration) and force detection of wearable sports equipment. It also holds significant application potential in human-machine interaction, intelligent robotic tactile sensing, flexible electronic skin, and other scenarios.

Besides, Hong et al. [33] developed a bionic flexible capacitive pressure sensor (BFCPS) with a unique mantis leg-inspired microstructured dielectric layer, using PDMS as the core material and conductive silver adhesive as the electrode material. The fabrication process involved preparing a polylactic acid (PLA) mold with a mantis leg bionic structure via 3D printing, followed by spin coating, curing, and self-assembly techniques. COMSOL Multiphysics 5.6 software was employed to construct two-dimensional (2D; size: 16 × 8.1 mm) and 3D (size: 16 × 10 × 8.5 mm) finite element models of the sensor's dielectric layer. Mechanical simulations were conducted by applying different displacement loads to the bionic microstructure (2D model: 1, 2, 4 mm; 3D model: 0.7, 1.4, 2.1 mm), yielding the stress-strain distribution law of the bionic structure. It was clarified that significant stress concentration occurs at the bent part of the dielectric layer under pressure, and this concentration

becomes more pronounced with increasing displacement – a key behavioral characteristic. The simulation results were mutually verified with experimental data, both confirming that the bionic structure can effectively enhance sensor sensitivity. The finally optimized sensor achieved a sensitivity of 0.836 /kPa in the range of 0–1 kPa, with response/recovery times of 40 ms/50 ms and a detection range spanning 12 Pa to 333 kPa. This sensor has been successfully applied in robot grasping object recognition (e.g., marker pens, oranges), human posture monitoring (e.g., cheek bulging, joint bending, respiration, and plantar pressure), and Morse code transmission (e.g., information such as “ABCDEF,” “Anhui,” and “Sensor”). It holds significant application potential in wearable electronic devices, intelligent robotic tactile sensing, human–machine interaction systems, and healthcare monitoring scenarios.

### **3.3 FEM design on flexible triboelectric tactile sensors**

As a rapidly emerging category of flexible tactile sensors, triboelectric tactile sensors – particularly those based on triboelectric nanogenerators (TENGs) – exhibit remarkable potential in wearable electronics, electronic skin, and Internet of Things (IoT) sensing systems owing to their self-powered operation, high sensitivity, and broad material adaptability. Their working principle is governed by the coupled effects of contact electrification and electrostatic induction, where external mechanical stimuli such as contact–separation, sliding, or deformation induce charge transfer across interfaces, thereby producing measurable electrical outputs. The electrical performance of TENG-based sensors, including output voltage, current, and sensitivity, is intricately influenced by factors such as material selection, surface microstructure, interfacial adhesion, and operational mode. To quantitatively explore these complex coupling effects, the finite element method (FEM) has emerged as a powerful analytical and predictive tool. Through the simulation of charge distribution, electric potential field evolution, and stress–strain interactions, FEM enables the elucidation of the electromechanical coupling mechanisms underlying triboelectric energy conversion, thus providing a robust theoretical foundation for structural optimization and performance enhancement.

#### *3.3.1 Microstructure design and surface engineering*

The surface microstructure is an effective means to enhance the output performance of the triboelectric sensor. By constructing structures at the micron or nanometer scale, the effective contact area can be significantly increased, the charge transfer efficiency can be enhanced, and thus the output voltage and current can be improved. FEM can be used to simulate the contact behavior, stress distribution, and potential field changes under different microscopic morphologies, guiding the selection of the optimal structural parameters.

For example, Das et al. [34] developed a triboelectric sensor (LAG-TEPS) utilizing laser-ablated graphene (LAG) and sandpaper-templated microstructured PDMS. The sensor demonstrates high sensitivity of 7.697 /kPa under low pressure (<1 kPa), a fast response time below 9.9 ms, and robust stability for over 4,000 cycles. Through finite element simulations in COMSOL, the team investigated the electric potential distribution across electrodes with varying spacings (0.5 and 1.0 mm;). The results reveal that the peak-to-peak output voltage ( $V_{pp}$ ) drops

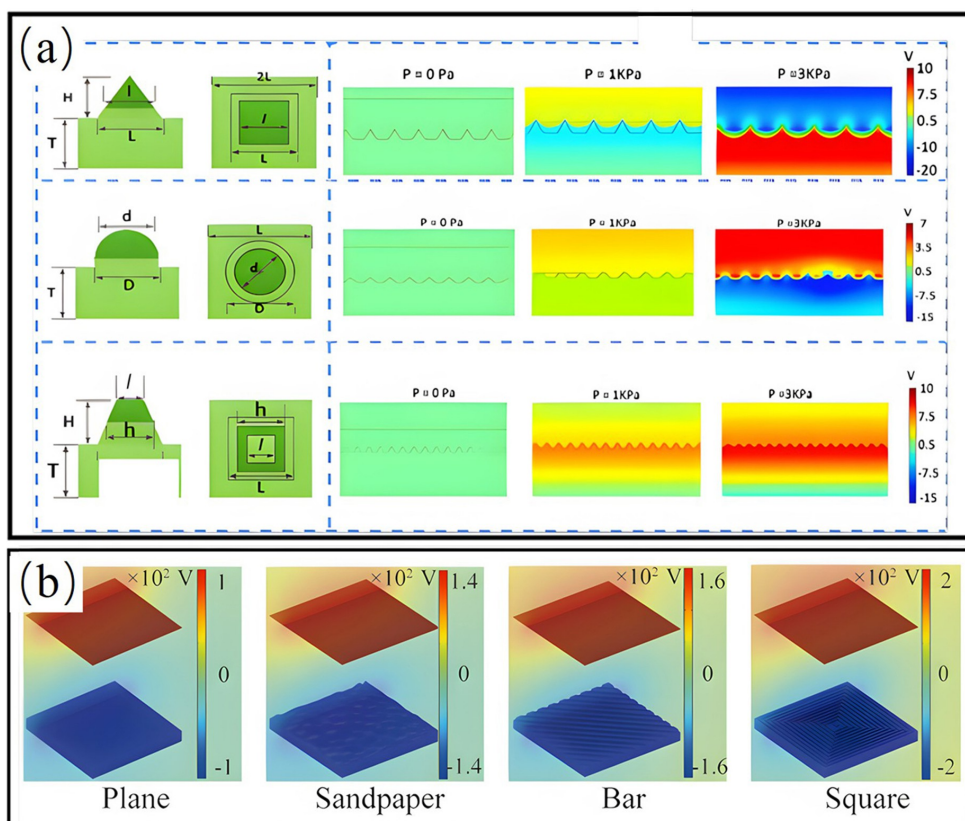


**Figure 5.** FEM on the microstructure design and surface engineering of triboelectric tactile sensors. (a) Simulated potential distribution and contact length variation under applied pressure for different microstructures (pyramid, hemisphere, and trapezoid) (used with permission [35]). (b) Simulated potential distribution of the square-loop microstructure at varying distances and corresponding output voltage comparison (used with permission [36]).

significantly as the electrode separation increases, thereby confirming an inverse relationship between them. This simulation work provides a theoretical foundation for the sensor's high sensitivity and rapid response, which are critical for its successful application in human gesture recognition and pulse monitoring.

In another work, Ahmed et al. [35] systematically investigated the influence of various microstructures (pyramidal, hemispherical, and trapezoidal) on the sensitivity of a stretchable triboelectric pressure sensor. Through COMSOL simulations, they analyzed the contact length, electric potential distribution, and amount of charge transfer ( $\Delta C/C_0$ ) for each structure under different pressure levels (0–3 kPa) (see **Figure 5a**). The results indicated that the hemispherical microstructure yielded the highest sensitivity and charge transfer efficiency under identical pressure, exhibiting the most uniform potential distribution and the most pronounced change in contact area. The simulations also revealed the influence of microstructure spacing and size on sensitivity: larger spacing led to more significant structural deformation and, consequently, higher sensitivity. This study provides a reliable simulation basis for the selection and geometric parameter optimization of microstructures in triboelectric tactile sensors.

Besides, Duan et al. [36] designed a surface featuring square-loop microstructures for their self-powered non-contact TENG sensor. They employed COMSOL simulations to compare the electric potential distributions of planar, sandpaper-textured, strip-patterned, and square-loop structures at varying distances (see **Figure 5b**). The simulation results demonstrated that the square-loop microstructure yielded the highest output voltage, attributable to its larger specific surface area and an enhanced electrostatic induction effect reminiscent of an antenna array. Furthermore, the study experimentally validated the simulation findings, showing that a square-loop structure with a  $7\text{ }\mu\text{m}$  wall width achieved a peak output voltage of 29 V. This agreement between simulation and experiment verifies the effectiveness of the FEM in optimizing microstructures.

### *3.3.2 Composite materials and multilayer design*

To enhance the dielectric properties and charge storage capacity of triboelectric tactile sensors, researchers often employ composite materials or multilayer stacked structures. FEM can be utilized to analyze the interfacial polarization behavior, electric field distribution, and the corresponding enhancement mechanisms for the output performance in such multilayer systems.

For example, Lee et al. [37] designed a multilayer TENG structure based on PDMS, which featured an alternating stack of ZTO@SMC (zinc stannate-decorated CNTs) and BTO (barium titanate) composite layers. The electric field distribution and potential output of single-, double-, triple-, and five-layer structures were analyzed via COMSOL simulations. The results revealed that the multilayer configuration significantly enhanced the interfacial polarization effect, leading to localized electric field concentration and improved charge transfer efficiency. Consequently, the five-layer structure achieved an output voltage of 797 V, a current of  $65\text{ }\mu\text{A}$ , and a power density of  $14.21\text{ W/m}^2$ . Furthermore, simulations indicated that the ZTO@SMC layers functioned as a conductive network to facilitate charge transport, while the BTO layers increased the dielectric constant and charge storage capacity. The synergistic effect between these two layers was identified as the key factor in enhancing the TENG's output performance.

In another work, Wang et al. [38] fabricated a porous triboelectric pressure sensor based on an MXene/PDMS composite film. They systematically tuned the mass ratio of MXene to PDMS to optimize both film porosity and electrical output. The potential distribution between an aluminum foil and the MXene/PDMS film at various separation distances was simulated using COMSOL. The simulations revealed that within a critical interface distance range of 0–0.10 mm, the surface potential on the aluminum foil changed at a high rate of  $50.51\text{ V/mm}$ . This simulation result provides a theoretical explanation for the sensor's high sensitivity ( $659.37\text{ mV/N}$ ) and rapid response time ( $<30\text{ ms}$ ) in the low-pressure regime.

Besides, triboelectric tactile sensors are prone to performance degradation due to mechanical fatigue or physical damage in practical applications. The incorporation of self-healing materials presents a viable strategy to significantly enhance their operational longevity and environmental adaptability. The FEM can be employed to simulate the mechanical restitution behavior and the retention of electrical performance throughout the self-healing process.

For example, Zhao et al. [39] pioneered the use of a self-healing hydrogel as a triboelectric material in a TENG, creating a flexible sensor capable of recovering from

cutting damage. The hydrogel exhibited excellent elastic recovery even at 200% tensile strain and retained 88% of its original voltage output after self-healing. The researchers employed COMSOL simulations to analyze the electric potential distribution during the contact-separation process between the hydrogel and a fluorinated ethylene propylene (FEP) film. The simulations confirmed that the healed hydrogel could maintain a stable electrical potential output, a finding consistent with experimental data. These results underscore the significant application potential of self-healing hydrogels in wearable triboelectric devices. Furthermore, the integration of triboelectric tactile sensors into wearable systems necessitates device attributes such as flexibility, conformability, and multi-channel sensing capability. The FEM can be employed to analyze key aspects for integration, including the sensor-skin coupling behavior, signal crosstalk between adjacent units, and the consistency of electrical output under complex mechanical deformations.

In another work, Bai et al. [40] developed a cylindrical rotary TENG to efficiently capture energy from rotational motion. COMSOL simulations visualizing the charge dynamics at the PTFE-metal sliding interface revealed an “in-plane polarization” mechanism and detailed charge transfer path. The strong agreement between these simulations and experimental electrical output data validates the design’s effectiveness for continuous energy harvesting from sources like wind and flowing water.

### **3.4 FEM design of flexible temperature sensors**

Flexible temperature sensors have emerged as essential components in wearable and intelligent monitoring systems, enabling continuous detection of body temperature, ambient fluctuations, and device thermal states [41]. Their lightweight structure, mechanical flexibility, and skin conformability offer significant advantages over traditional rigid thermometers, supporting long-term, real-time temperature monitoring without discomfort or signal drift [42]. To enhance the sensitivity, thermal stability, and mechanical robustness, researchers increasingly employ FEM to elucidate the coupled thermal-mechanical behavior of these devices. Typically, flexible temperature sensors operate through resistive, thermoelectric, or thermochromic mechanisms. Among these, resistive-type sensors – based on the temperature-dependent variation of electrical resistance – are the most widely utilized due to their simple structure and stable output. Thermoelectric sensors, by contrast, rely on the Seebeck effect between dissimilar materials, while thermochromic systems visualize temperature changes via color transition. Structurally, most flexible temperature sensors adopt a multilayer configuration composed of a deformable substrate, a thermosensitive conductive layer, and an encapsulation film. Common substrates, such as polyimide (PI), polyethylene terephthalate (PET), or PDMS, ensure flexibility and biocompatibility, whereas conductive layers integrating graphene, CNTs, or MXenes provide high thermal-electrical responsiveness [43, 44].

For example, Sreevanya et al. [45] focused on the electrode architecture of temperature sensors for wearable electronics. They proposed a silver-electrode design patterned with periodic “kirigami” cuts and built a finite-element model in ANSYS to simulate its deformation and resistance evolution under large strains (up to 50%). Systematic simulations were carried out to evaluate how different kirigami motifs influence the mechanical robustness and electrical stability of the silver film. The results show that, compared with conventional stretchable conductors, the kirigami electrode maintains stable electrical performance even under extreme

elongation, and the predictions agree closely with experimental data, underscoring the great potential of kirigami architectures for flexible electronic devices.

In another work, Lu et al. [46] reported a fiber-type flexible sensor based on a graphene/CNT composite that is highly deformable and exhibits an exceptionally large temperature-coefficient of resistance (TCR) of  $-20.2\%/^{\circ}\text{C}$  over the clinically relevant window of  $-15^{\circ}\text{C}$  to  $60^{\circ}\text{C}$ , making it ideal for on-body, dynamic thermometry. A key innovation of the work is the systematic integration of finite-element modeling: the electro-thermal behavior of the fiber under combined thermal and mechanical loads was simulated in COMSOL Multiphysics (see **Figure 6a**). The simulations reveal that conductive pathways within the composite evolve non-linearly when temperature and strain vary simultaneously, governing both resistance change and sensitivity. By comparing FEM predictions with experimental data acquired during cyclic bending and heating, the authors validated the model's accuracy and robustness under realistic wear conditions, achieving a prediction error of  $<10\%$ .

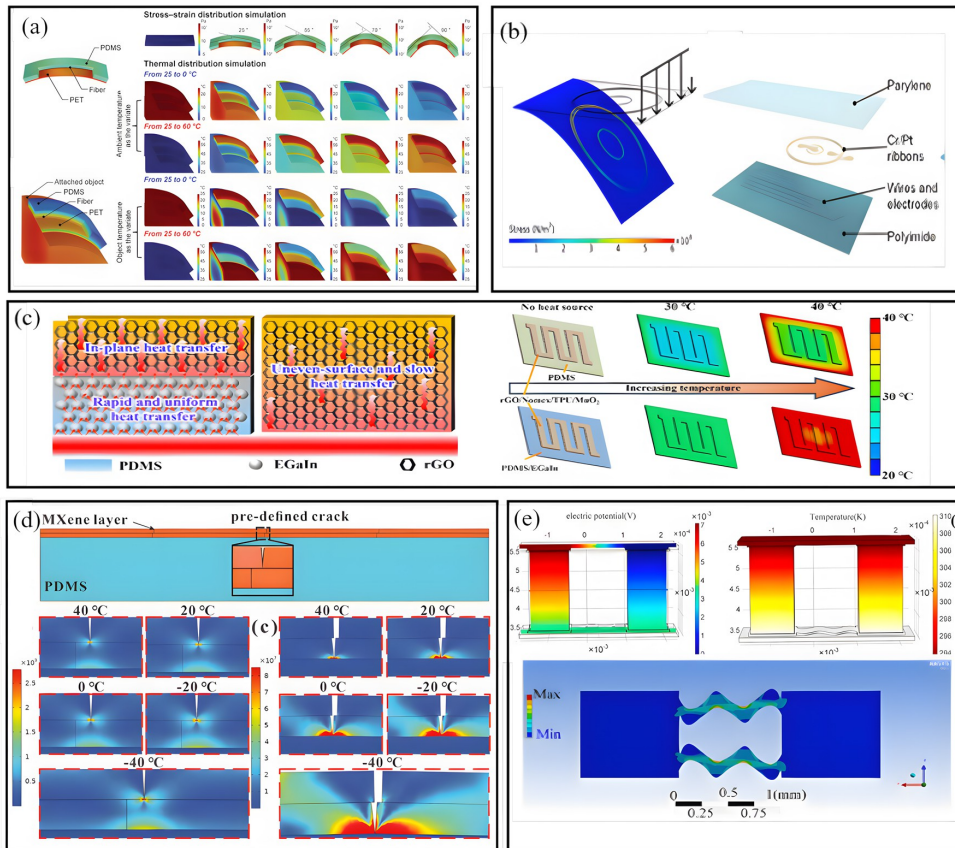


**Figure 6.**

FEM on structured temperature sensors. (a) FEM of the flexible temperature sensor (used with permission [46]). (b) The stress in two ribbons of the sensor under bending simulated by FEM using COMSOL (used with permission [47]). (c) Comparative ANSYS thermal analysis between the interface layer with and without liquid metal (used with permission [48]). (d) Between MXene layers and at MXene/PDMS interface at temperature ranging from 40 to  $-40^{\circ}\text{C}$ , respectively (used with permission [49]). (e) Simulation of the electric field distribution across the single thermocouple (used with permission [49]).

In addition, Wang et al. [47] proposed a thermo-aware compensation strategy and fabricated a new device architecture with active error-correction capability. A dedicated temperature-sensing channel is monolithically integrated to continuously acquire thermal information; these data are fed back in real time to decouple strain-induced resistance changes from thermally driven drifts, thereby producing a strain-only output. To reveal the underlying thermo-mechanical coupling, the authors built finite-element models that reproduce temperature profiles and stress maps under combined thermal (10°C–60°C) and tensile loads (see **Figure 6b**). Simulated heat-flow paths and stress-concentration sites agree closely with infrared and strain-field measurements, confirming that the coupled mechanism must be embedded in the design loop. Experimental benchmarking shows that compensated sensors maintain a stable, linear response across the entire temperature and strain ranges, exhibiting markedly improved repeatability compared with uncompensated counterparts.

Besides, Wang et al. [48] focused on the core design philosophy of temperature sensors and highlighted the critical role of FEA in structural optimization, thermo-mechanical coupling modeling, and performance prediction. Using COMSOL Multiphysics, they developed a multiphysics-coupled model to numerically simulate the thermal field distribution and stress concentration of the device under various temperature gradients, tensile deformations, or bending states (see **Figure 6c**). The simulation results revealed that key factors influencing device performance include the thermal conductivity of the temperature-sensitive material, the mechanical modulus of the flexible substrate, and the interfacial adhesion properties. By integrating FEM into the design workflow of flexible temperature sensors, researchers can predict thermal response characteristics, stress evolution patterns, and the impact of deformation before device fabrication, significantly enhancing development efficiency and device reliability.

Furthermore, Song et al. [49] developed a flexible temperature sensor based on an MXene/PDMS composite and systematically investigated its bending-fatigue behavior at sub-zero temperatures (down to –20°C). The device was fabricated by lamination into a sandwich architecture: MXene nanosheets dispersed in the PDMS matrix serve as the compliant, conductive sensing layer, yielding high initial sensitivity and mechanical flexibility. Experimentally, the sensor was subjected to hundreds of bending cycles at various temperatures and radii while its capacitance was monitored. The data revealed accelerated performance degradation at low temperatures, attributed to intensified stress concentration. To uncover the underlying micro-mechanism, the authors performed FEM to map the internal stress distribution and crack-evolution path during bending (see **Figure 6d**). A 2D multilayer model built in COMSOL Multiphysics faithfully reproduced the strain-localization phenomenon. Simulations showed that the elastic modulus of PDMS increases at low temperatures, making the MXene conductive network more susceptible to fracture and thus causing the observed electrical decay. The close agreement between simulation and experiment validates the use of FEM for predicting failure modes and guiding structural optimization. This work not only clarifies the low-temperature fatigue characteristics of MXene/PDMS flexible sensors but also demonstrates the broader potential of finite-element modeling for reliability studies of flexible electronics operating in extreme environments.

Lastly, Feng et al. [50] adopted a Bi<sub>2</sub>Te<sub>3</sub>-based composite as the thermoelectric active layer and improved both the Seebeck coefficient and power output by optimizing the device architecture and thickness matching. On this basis, the team

performed finite-element simulations to systematically map the temperature field and power response under varying heat-flux densities, contact areas, and ambient temperature differences. The simulated results agreed closely with experimental data, confirming that deliberate thermal-management design is pivotal for performance enhancement. By integrating FEM thermal modeling with optimization of the material's thermoelectric parameters, the study provides both theoretical guidance and an engineering demonstration for high-performance flexible thermoelectric temperature sensors, highlighting the value of multiphysics coupling in sensor design (see **Figure 6e**). The fabricated device can continuously deliver microwatt-level power – sufficient to drive low-power wireless sensor nodes – offering promising prospects for health monitoring, smart garments, and remote temperature-sensing systems.

#### **4. Conclusion(s)**

FEM has firmly established itself as a transformative tool in the design and development of flexible tactile sensors, effectively addressing the critical limitations of traditional trial-and-error approaches, which are often plagued by low efficiency, high costs, and insufficient mechanistic insight. This chapter has systematically delineated the pivotal role of FEM across various sensor types – including resistive, capacitive, triboelectric, and temperature sensors – highlighting its three principal advantages: (1) serving as a virtual design platform that accelerates the optimization of key parameters (e.g., microcrack spacing, porosity, electrode geometry), potentially reducing experimental iterations by up to 60%; (2) acting as a cost-saving strategy by predicting performance outcomes, thereby minimizing the need for physical prototypes and associated costs by 30%–50%; (3) providing unprecedented mechanistic insight by visualizing internal field distributions (stress, current density, electric potential), which elucidates fundamental sensing mechanisms such as the coupling between stress concentration and the evolution of conductive pathways in multilayer and bioinspired architectures. Specifically, FEM-guided optimization has been instrumental in achieving remarkable advancements. It has enabled the realization of highly sensitive and stable resistive sensors based on wrinkled graphene and porous aerogels; expanded the linear sensing range of capacitive sensors by precisely modeling electrode-dielectric contact behavior; revealed the critical relationships between charge distribution, morphological features, and output performance in TENGs, paving the way for efficient self-powered systems; and enhanced the response speed and interference immunity of temperature sensors through coupled thermo-mechanical simulations. Collectively, the integration of FEM has not only propelled device performance to new heights but also laid a solid foundation for their application in wearable health monitoring, soft robotics, and advanced human-machine interaction.

#### **Acknowledgments**

This work was supported by the Fund by Science Research Project of Hebei Education Department (No. JZX2024024) and the Research Startup Fund by Zhejiang Sci-Tech University (No. 24212186-Y).

## Conflict of Interest

The authors declare no conflict of interest.

## Author details

Peng Wang<sup>1,2</sup>, Yu Lv<sup>1</sup>, Li Chen<sup>1</sup>, Xiufeng Li<sup>1</sup>, Xiaoyu Cao<sup>1</sup> and Chuizhou Meng<sup>1\*</sup>

1 Research Institute of Wearable Electronic Materials & Devices, School of Materials Science & Engineering, Zhejiang Sci-Tech University, Hangzhou, China

2 School of Mechanical Engineering, University of Jinan, Jinan, China

\*Address all correspondence to: [mcz@zstu.edu.cn](mailto:mcz@zstu.edu.cn)

## IntechOpen

---

© 2025 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Pyo S, Lee J, Bae K, et al. Recent progress in flexible tactile sensors for human-interactive systems: from sensors to advanced applications. *Advanced Materials*. 2021;**33**(47):2005902. DOI: 10.1002/adma.202005902
- [2] Liu Y, Wang W, Chen X. Progress and prospects in flexible tactile sensors. *Frontiers in Bioengineering and Biotechnology*. 2023;**11**:1264563. DOI: 10.3389/fbioe.2023.1264563
- [3] Cui X, Zhang W, Lv M, et al. Flexible and wearable tactile sensors for intelligent interfaces. *Materials*. 2025;**18**(17):4010. DOI: 10.3390/ma18174010
- [4] Kumar KS, Chen PY, Ren H. A review of printable flexible and stretchable tactile sensors. *Research*. 2019;**2019**:3018568. DOI: 10.34133/2019/3018568
- [5] Wei Y, Li B, Domingos M, et al. Fully 3D printed flexible, conformal and multi-directional tactile sensor with integrated biomimetic and auxetic structure. *Communications Engineering*. 2023;**2**(1):80. DOI: 10.1038/s44172-023-00131-x
- [6] Chang WY, Fang TH, Yeh SH, et al. Flexible electronics sensors for tactile multi-touching. *Sensors*. 2009;**9**(2):1188–1203. DOI: 10.3390/s9021188
- [7] Zhang R, Lin JT, Popa DO. Finite element analysis of a flexible tactile sensor with circular pattern. In: *Proceedings of the 2022 IEEE International Conference on Flexible and Printable Sensors and Systems (FLEPS)*; 2022; Vienna, Austria. DOI: 10.1109/FLEPS53764.2022.162
- [8] Li T, Liu L, Zhang D, et al. Recent progress on flexible tactile sensors: From microstructural design to intelligent sensing applications. *Materials Today Physics*. 2025;101793. DOI: 10.1016/j.mtphys.2025.101793
- [9] Yeo JC, Yu J, Koh ZM, et al. Wearable tactile sensor based on flexible microfluidics. *Lab on a Chip*. 2016;**16**(17):3244–3250. DOI: 10.1039/c6lc00579a
- [10] Ren H, Li W, Ding Y, et al. Flexible tactile sensors for enhancing robotic perception. *Smart Wearable Technology*. 2025;1:A2. DOI: 10.47852/bonviewSWT52025770
- [11] Jiao C, Wang C, Wang M, et al. Finite element analysis model of electronic skin based on surface acoustic wave sensor. *Nanomaterials*. 2023;**13**(3):465. DOI: 10.3390/nano13030465
- [12] Li M, Torah R, Liu J, et al. Finite element analysis (FEA) modelling and experimental verification to optimise flexible electronic packaging for e-textiles. *Microsystem Technologies*. 2022;**28**(6):1515–1524. DOI: 10.1007/s00542-020-04813-w
- [13] Wang Y, Chen J, Mei D. 2019. Flexible tactile sensor array for slippage and grooved surface recognition in sliding movement. *Micromachines*. **10**(9):579. DOI:10.3390/mi10090579
- [14] Kumar KS, Chen PY, Ren H. A review of printable flexible and stretchable tactile sensors. *Research*. 2019;**2019**:3018568. DOI: 10.34133/2019/3018568
- [15] Meng L, Ding L, Khan AM, et al. Mathematical modeling of flexible

- printed circuit configuration: A study in deformation and optimization. *Scientific Reports*. 2024;**14**:14128. DOI: 10.1038/s41598-024-64770-6
- [16] Tang X, Yang W, Yin S, et al. Controllable graphene wrinkle for a high-performance flexible pressure sensor. *ACS Applied Materials & Interfaces*. 2021;**13**:20448–20458. DOI: 10.1021/acsami.0c22784.
- [17] Wei L, Wu Z, Tang S, et al. Tracheid-inspired nanoarchitected carbon-based aerogels with ultra-compressibility for wearable piezoresistive sensors. *Carbon*. 2023;**203**:386–396. DOI: 10.1016/j.carbon.2022.11.081
- [18] Cho J, Kim DH, Choi YJ, et al. Synergistic micro-crack engineering and auxetic metamaterials for enhanced sensitivity in stretchable strain sensors. *Advanced Materials Technologies*. 2025; e01027. DOI: 10.1002/admt.2025001027
- [19] Han HJ, Jeong JW, Yang SR, et al. Nanotransplantation printing of crystallographic-orientation-controlled single-crystalline nanowire arrays on diverse surfaces. *Communications Engineering*. 2017;**11**:11642–11652. DOI: 10.1021/acsnano.7b06696
- [20] Takamatsu S, Takakuwa M, Fukuda K, et al. Flexible no-drift data glove using ultrathin silicon for the metaverse. *Device*. 2025;**3**:100686. DOI: 10.1016/j.device.2024.100686
- [21] Gupta U, Lau JL, Chia PZ, et al. All knitted and integrated soft wearable of high stretchability and sensitivity for continuous monitoring of human joint motion. *Advanced Healthcare Materials*. 2024; **16**:2202987. DOI: 10.1002/adhmi.202202987
- [22] Wang N, Sun S, Lei T, et al. Enhanced piezoresistive sensors and triboelectric nanogenerators based on multifunctional wood-based aerogels for smart home systems. *Advanced Functional Materials*. 2025;e21776. DOI: 10.1002/adfm.202521776
- [23] Wang B, Li Y, Ma L, et al. In situ rapid gelation and osmotic dehydration-assisted preparation of graphene aerogel and its application in piezoresistive sensors. *Journal of Colloid and Interface Science*. 2025;**138906**. DOI: 10.1016/j.jcis.2025.138906
- [24] Wang Z, Guan X, Huang H, et al. Full 3D printing of stretchable piezoresistive sensor with hierarchical porosity and multimodulus architecture. *Advanced Functional Materials*. 2019;**29**:1807569. DOI: 10.1002/adfm.201807569
- [25] Yan J, Hou S, Wu S, et al. Micro/nanosphere-regulated MXene-based flexible piezoresistive sensor with ultra-high sensitivity and broad detection range. *Small*. 2025;e08394. DOI: 10.1002/smll.202508394
- [26] Zhang X, Sun K, Ma J, et al. Gradient porous flexible pressure sensors with the relay effect for high-accuracy braille-to-speech recognition. *ACS Sensors*. 2025;**10**:7035–7050. DOI: 10.1021/acssensors.5c02096
- [27] Gabbi C, Peruzzini M, Salerno M, et al. FEM-guided geometric parameterization of paper-based piezoresistive sensors for low-cost flexible electronics. *Sensors and Actuators A: Physical*. 2024;**365**:115654. DOI: 10.1016/j.sna.2024.115654

- [28] Mishra RB, El-Atab N, Hussain AM, et al. Recent progress on flexible capacitive pressure sensors: From design and materials to applications. *Advanced Materials Technologies*. 2021;**6**:2001023. DOI: 10.1002/admt.202001023
- [29] Yang J, Luo S, Zhou X, et al. Flexible, tunable, and ultrasensitive capacitive pressure sensor with microconformal graphene electrodes. *ACS Applied Materials & Interfaces*. 2019;**11**:14997–15006. DOI: 10.1021/acsami.9b02049
- [30] Zhong Y, Gu FC, Wu LG, et al. Porous conductive electrode for highly sensitive flexible capacitive pressure sensor over a wide range. *Journal of Alloys and Compounds*. 2023;**934**: 167919. DOI: 10.1016/j.jallcom.2022.167919
- [31] Zhang ZA, Gui XC, Hu QM, et al. Highly sensitive capacitive pressure sensor based on a micropyramid array for health and motion monitoring. *Advanced Electronic Materials*. 2021;**7**:2100174. DOI: 10.1002/aelm.202100174
- [32] Guo XH, Zhao JJ, Hu B, et al. Flexible pressure sensor with high sensitivity and fast response based on bionic honeycomb-structured polydimethylsiloxane/aluminum oxide composites dielectric via 3-D printing. *IEEE Transactions on Electron Devices*. 2024;**71**:4283–4291. DOI: 10.1109/TED.2024.3401653
- [33] Hong WQ, Guo XH, Hu B, et al. Mantis leg-inspired flexible capacitive pressure sensor with high sensitivity and fast response for wearable devices and human-machine interaction. *IEEE Sensors Journal*. 2024;**24**:19698–19705. DOI: 10.1109/JSEN.2024.3392878
- [34] Das PS, Chhetry A, Maharjan P, et al. A laser ablated graphene-based flexible self-powered pressure sensor for human gestures and finger pulse monitoring. *Nano Research*. 2019;**12**:1789–1795. DOI: 10.1007/s12274-019-2433-5
- [35] Ahmed A, Hassan I, Zu J. Design guidelines of stretchable pressure sensors-based triboelectrification. *Advanced Engineering Materials*. 2018;**20**:1700997. DOI: 10.1002/adem.201700997
- [36] Duan J, Yue G, Li H, et al. Self-powered noncontact triboelectric nanogenerators with microstructured square-loop surface and dielectric electron-blocking layer for far-distance motion perception and trajectory tracking. *Nano Energy*. 2024;**128**:109860. DOI: 10.1016/j.nanoen.2024.109860
- [37] Lee K, Ryu JH, Kim B, et al. Rational design of materials and multilayered structures in triboelectric nanogenerators to enhance energy-harvesting efficiency. *Chemical Engineering Journal*. 2025;**520**:166062. DOI: 10.1016/j.cej.2025.166062
- [38] Wang Z, Xu Z, Li N, et al. Flexible pressure sensors based on MXene/PDMS porous films. *Advanced Materials Technologies*. 2023;**8**:2200826. DOI: 10.1002/admt.202200826
- [39] Zhao K, Lv H, Meng J, et al. Triboelectrification-induced electricity in self-healing hydrogel for mechanical energy harvesting and ultra-sensitive pressure monitoring. *ACS Omega*. 2022;**7**:18816–18825. DOI: 10.1021/acsomega.2c01743
- [40] Bai P, Zhu G, Liu Y, et al. Cylindrical rotating triboelectric

nanogenerator. *ACS Nano*. 2013;7: 6361–6366. DOI: 10.1021/nn402491y

[41] Lu YF, Zhang HJ, Huang W, et al. Flexible temperature-pressure dual sensor based on 3D spiral thermoelectric Bi<sub>2</sub>Te<sub>3</sub> films. *Nature Communications*. 2024;15:46836. DOI: 10.1038/s41467-024-46836-1

[42] Krishnan A, Thiyagarajan K, Kodagoda S, et al. Wearable flexible temperature sensor suite for thermal-tactile perception. *IEEE Sensors Journal*. 2024;24: 39736–39743. DOI: 10.1109/JSEN.2024.3474705

[43] Chen S, Han X, Hong P, et al. A flexible temperature sensor for noncontact human-machine interaction. *Materials*. 2021;14:7112. DOI: 10.3390/ma14237112

[44] Sibinski M, Jakubowska M, Sloma M. Flexible temperature sensors on fibers. *Sensors*. 2010;10:7934–7946. DOI: 10.3390/s100907934

[45] Sreevanya GV, Lee SJ, Cheon H, et al. Highly stretchable and flexible kirigami patterned silver electrodes for wearable electronics. *Sensors and Actuators A: Physical*. 2024;378:115813. DOI: 10.1016/j.sna.2024.115813

[46] Lu Y, Zhang H, Zhao Y, et al. Robust fiber-shaped flexible temperature sensors for safety monitoring with ultrahigh sensitivity. *Advanced Materials*. 2024;36:2310613. DOI: 10.1002/adma.202310613

[47] Wang L, Zhu R, Li G. Temperature and strain compensation for flexible sensors based on thermosensation. *Advanced Materials*. 2019;12:1953–1961. DOI: 10.1021/acsami.9b21474

[48] Wang P, Sun GF, Meng CZ, et al. A flexible nanoparticle structured temperature sensor with high reproducibility and ultrafast response speed enabled by thermal expansion coefficient regulation and thermal conductivity enhancement. *Nano Materials Science*. 2025. DOI: 10.1016/j.nanoms.2025.10.007

[49] Song J-Y, Yu P-L, Li J-C. Low-temperature bending fatigue of MXene/PDMS flexible pressure sensor. *Advanced Materials Interfaces*. 2022;9:2201463. DOI: 10.1002/admi.202201463

[50] Feng R, Tang F, Zhang N, et al. Flexible, high-power density, wearable thermoelectric nanogenerator and self-powered temperature sensor. *ACS Applied Materials & Interfaces*. 2019;11:38616–38624. DOI: 10.1021/acsami.9b11435



# Uncertainty Quantification in Electromagnetic Computation

*Geoffrey Lossa and Olivier Deblecker*

## Abstract

In this work, uncertainty quantification is integrated to assess the sensitivity of a finite element (FE)-based numerical model to uncertainties in its input parameters. We compute the sensitivity of wound inductors' RL parameters extracted from the FE method with respect to geometric and material uncertainties across a wide range of frequencies (from DC to 1 MHz). To that end, we propose a 2D modeling of geometric uncertainties associated with conductors' positions in the winding window, and we compute the *Sobol* indices associated with a polynomial chaos expansion-based surrogate model of the uncertain FE model using Bayesian inference techniques.

**Keywords:** finite element, homogenization technique, uncertainty quantification, inference technique, polynomial chaos expansion, wound inductors

## 1. Introduction

Numerical models of a physical system almost always depend on uncertain parameters. For instance, wound inductors are characterized by material uncertainties (material properties that are not accurately known) and geometric uncertainties (2D positions of conductors in the winding window, as depicted in **Figure 1**, air-gap dimensions, etc.). Correct modeling of such uncertainties, as well as a good understanding of the sensitivity of physical quantities with respect to uncertainties, is fundamental for ensuring an optimal design of the component.

The finite element (FE) method can be employed to capture these uncertainties in RLC models using a Monte Carlo (MC) simulation approach [1]. However, performing a sensitivity analysis with such a model is very CPU-intensive. In this work, we develop a stochastic surrogate model of wound inductors based on polynomial chaos expansion (PCE) in order to mitigate this problem. We then compute the *Sobol* indices [2, 3] over a wide range of frequencies (from DC to 1 MHz) to quantify the sensitivity of RL parameters from the PCE surrogate model.

Sensitivity analysis is not limited to determining how the variability of the model is affected by the variability of each random input, but it can also be used to highlight unimportant inputs and thereby help reduce the dimension of the stochastic problem. In general, it can be based on the correlation between input and output parameters (correlation based on measurements), on values of partial derivatives of

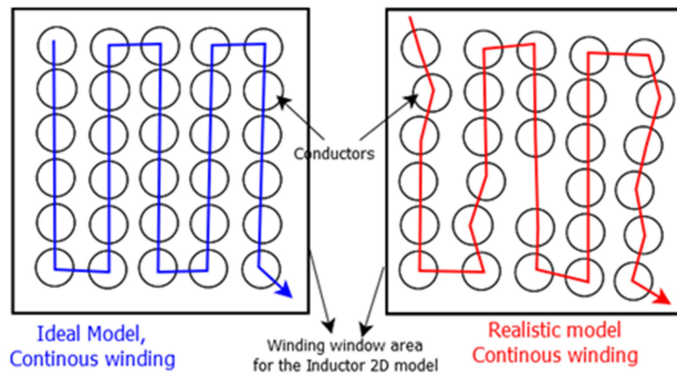


**Figure 1.**  
Illustration of geometric uncertainties in a 2D pattern of a five-layer winding.

the model at a given point (linear method), or even directly on the distribution of the model output. The global sensitivity analysis is based more specifically on the decomposition of the variance of the model response in terms of the contribution of each input parameter or their combinations. This is the case, for example, with the *Sobol* sensitivity indices [2].

## 2. Methodology

The uncertainties taken into account in this study arise from the random positions of conductors in the winding window and from the material properties of the ferrite core (magnetic permeability  $\mu_{core}$  and electrical conductivity  $\sigma_{core}$ ). The complex form of material properties allows for the natural consideration of electrical and magnetic losses in FE formulations.

Due to the high number of MC iterations with an embedded time-consuming FE-based model, especially with regard to the high number of geometric uncertainties, the reduction in the dimension of the problem is required and is achieved through the transformation of geometrical uncertainties (2D positions of conductors) into material ones (equivalent magnetic reluctivity  $\nu_{prox}$  related to the proximity effect and equivalent impedance  $Z_{skin}$  related to the skin effect) via a homogenization technique [4–6].

The overall methodology is built upon a stochastic surrogate model [4] for the computation of RL parameters, in combination with an original algorithm mimicking the manual winding operation [7] in order to model the uncertain positions of conductors in the winding window. It is depicted in **Figure 2**. We compute a PCE surrogate in order to reduce the computation time associated with the FE-based numerical model during calls needed to perform a MC simulation (with a brute FE model). In addition to this, we compute the *Sobol* coefficients from the PCE surrogate, which does not require any additional evaluation of the deterministic model (which is expensive in terms of resources and computation time) once it has been built. This is due to one of the important properties of PCE: that of containing information on its ANOVA (analysis of variance) decomposition.

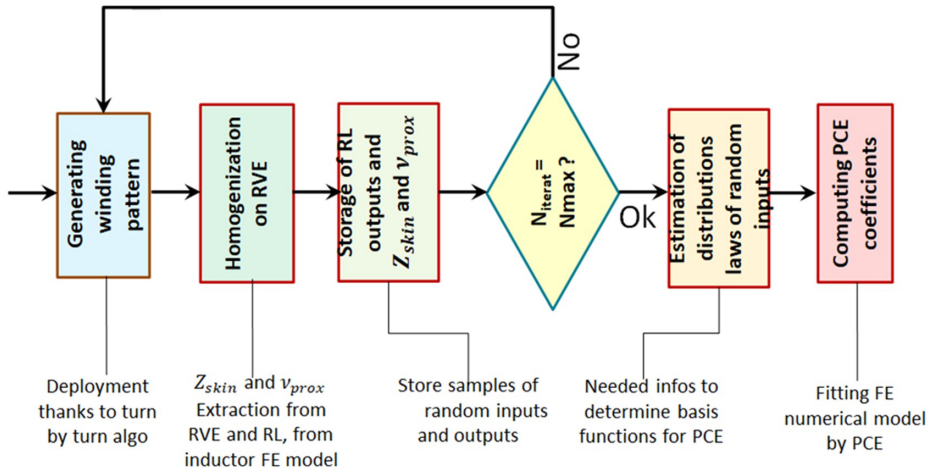


**Figure 2.**  
 Process of PCE surrogate construction.

According to this chart,  $N_{iter}$  is the number of iterations needed to build the PCE surrogate from distributions of the reduced random input vector.  $N_{iter}$  is bounded by  $N_{max}$  which is a value linked to the dimension of the random input vector. In our case, without the homogenization process, which allows us to transform geometrical uncertainties into material ones, this dimension can easily reach an order of magnitude close to 100.

A 2D magneto-harmonic formulation with massive conductors and appropriate limit conditions on an optimal representative volume element (RVE) [8] in the winding allows the extraction of the desired equivalent properties (electric conductivity and magnetic reluctivity, for instance). Finally, the dimension of the random input is reduced to seven components, among which we find:

- $X_1 = \Re(Z_{skin})$ , image of losses due to skin effect phenomena in the winding;
- $X_2 = \Im(Z_{skin})$ , image of the reactive power transmitted by the winding;
- $X_3 = \Re(v_{prox})$ , image of the energy associated with the proximity effect in the winding;
- $X_4 = \Im(v_{prox})$ , image of losses linked to the proximity effect between conductors;
- $X_5 = \Re(v_{core})$ , image of the magnetic energy in the ferrite core;
- $X_6 = \Im(v_{core})$ , image of the ferrite core losses;
- $X_7 = \sigma_{core}$ , image of losses due to the eddy current in the ferrite core. It is a scalar depending on the imaginary part of the electric permittivity  $\epsilon_{core}$ .

For the computation of the PCE surrogate model, we need to know the distribution law of the obtained reduced random input vector. This problem is

associated with the inverse problem, where unknown parameters (related to distributions) are estimated based on experimental data, which are indirectly associated with these parameters through a computational model. This problem can be solved using Bayesian methods, particularly the Bayesian inference technique [9]. This technique consists of the backward propagation of information about observations in order to extract the distribution of the model inputs. For instance, in statistical inference, one considers that the data consist of independent realizations of an underlying random vector and an assumption on the shape of the probability density function PDF (e.g., Weibull, Gaussian, log-normal, etc.). The tools used here are based on UQLab, which is a framework that can be plugged into MATLAB to perform uncertainty quantification [10].

The *Sobol* decomposition of a truncated PCE can be established analytically in terms of increasing orders as follows.

$$\begin{aligned}\mathcal{M}_p(X) = & a_0 + \sum_{i=1}^M \sum_{a \in \mathcal{I}_i} a_a \Psi_a(X_i) \\ & + \sum_{1 \leq i_1 < i_2 \leq M} \sum_{a \in \mathcal{I}_{i_1, i_2}} a_a \Psi_a(X_{i_1}, X_{i_2}) + \dots \\ & + \sum_{1 \leq i_1 < \dots < i_s \leq M} \sum_{a \in \mathcal{I}_{i_1, \dots, i_s}} a_a \Psi_a(X_{i_1}, \dots, X_{i_s}) + \dots \\ & + \sum_{a \in \mathcal{I}_{1, \dots, M}} a_a \Psi_a(X)\end{aligned}\quad (1)$$

where the set  $\mathcal{I}_{i_1, \dots, i_s}$  corresponds to  $\Psi_a$  polynomials depending on all parameters  $\{X_{i_1}, \dots, X_{i_s}\}$  and nothing but them. It is defined as:

$$\begin{aligned}\mathcal{I}_{i_1, \dots, i_s} = & \{\alpha \in \mathbb{N}^M : 0 \leq |\alpha| \leq p, \\ & \forall k \in \{1, \dots, M\} / \{i_1, \dots, i_s\}, \alpha_k = 0\}\end{aligned}\quad (2)$$

One can highlight the dependence of each multivariate polynomial on each subset of input parameters for clarity. This allows us to deduce the variance of the decomposition in the following way:

$$\mathbb{V}[\mathcal{M}(X)] \equiv D_p = \sum_{i=1}^M D_i + \sum_{1 \leq i < j \leq M} D_{ij} + \dots + D_{1, \dots, M} \quad (3)$$

where  $D_{i_1, \dots, i_s}$  refer to partial variances and are defined as:

$$D_{i_1, \dots, i_s} = \mathbb{V}[\mathcal{M}_{i_1, \dots, i_s}(X_{i_1}, \dots, X_{i_s})], \quad s = 1, \dots, M \quad (4)$$

From these last expressions, it is possible to define the sensitivity index  $\mathcal{S}_{i_1, \dots, i_s}^p$  of the model response for the subset of random input variables  $\{X_{i_1}, \dots, X_{i_s}\}$  as:

$$\mathcal{S}_{i_1, \dots, i_s}^p = \frac{D_{i_1, \dots, i_s}}{D_p} = \frac{1}{D_p} \sum_{a \in \mathcal{I}_{i_1, \dots, i_s}} a_a^2 \quad (5)$$

In addition, total sensitivity indices, also based on PCE, can be defined to assess the total effect of an input parameter as follows:

$$S_i^{T,p} \sum_{i_1, \dots, i_s \in \mathcal{I}_i^+} \frac{D_{i_1, \dots, i_s}}{D_p} \frac{1}{D_p} \sum_{a \in \mathcal{I}_i^+} a_a^2 \quad (6)$$

where  $\mathcal{I}_i^+$  indicates the set of all indices involving the  $i^{th}$  input parameter with a non-zero index, that is:

$$\mathcal{I}_i^+ \equiv \{\alpha \in \mathbb{N}^M : 0 \leq |\alpha| \leq p, \alpha_i \neq 0\} \quad (7)$$

## 2.1. Algorithm for generating realistic 2D winding patterns for stochastic FE modeling of wound inductors

According to this algorithm, conductors are deployed turn after turn. It aims to emulate the manual process as it would be carried out by an operator (**Figure 3**). The algorithm includes the following actions.

- The first layer conductors are the ones that are deployed first. For these, we only observe uncertainties  $\delta_y$  according to their vertical axis because they are placed against the magnetic core's central column.
- Conductors of the second and subsequent layers, except for their last conductors, are deployed to intersect with conductors of previous layers while maintaining two points of tangency with the latter.

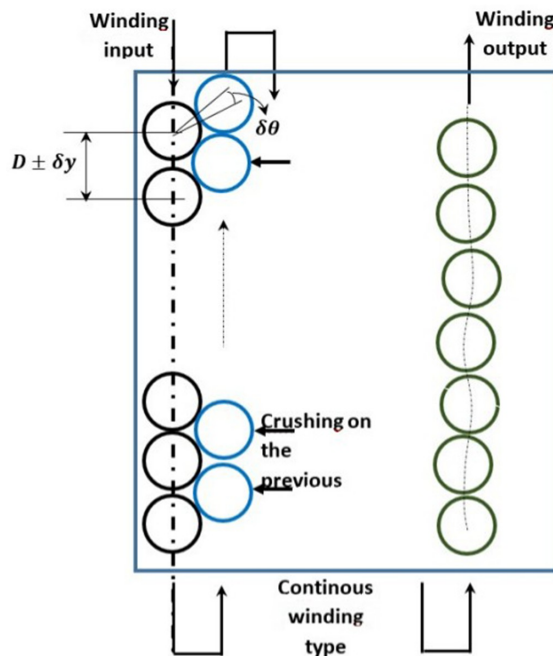


**Figure 3.**  
 Principle of the simulation of winding construction.

- In case there is an overlap between two conductors, it is necessary to redeploy the conductor corresponding to the turn being deployed in order to maintain two points of tangency: one with the conductor of the same layer as itself.
- The last conductors of the second and subsequent layers are positioned to present an uncertainty of  $\delta_\theta$  angle and a point of tangency with their corresponding conductors in the previous layer.
- In the event of an overlap with the conductor of the same layer, a redeployment is carried out to contract two points of tangency (with the conductor of the same layer and its corresponding conductor from the previous layer).

This original method of deploying the conductors in the winding window can be used to support the propagation of geometric uncertainties in a deterministic model, such as the FE method embedded into a MC simulation. To validate this approach, let us analyze the influence of geometric uncertainties related to the positions of conductors in the extraction of RLC parameters of a wound inductor made of an MnZn type ferrite core. Its magnetic core, without an air-gap, is considered ideal – that is, without electrical and magnetic losses (a relative magnetic permeability of 1,000 and 0 electric conductivity are considered). For this analysis, a 2D model is sufficient due to the axisymmetric characteristic of the geometry (see **Figure 4**). The copper conductors have a diameter of 0.315 mm. They are deployed in the winding window using our developed algorithm simulating the winding process.

Results presented in **Figure 5** are from a MC simulation of 390 iterations. According to this figure, one can observe an increasing dispersion of the  $R$  parameter due to the appearance of skin and proximity effects with increasing frequency, while the  $L$  and  $C$  parameters maintain a quasi-constant dispersion over the analysis frequency range and decrease with increasing frequency. In this figure, dispersions

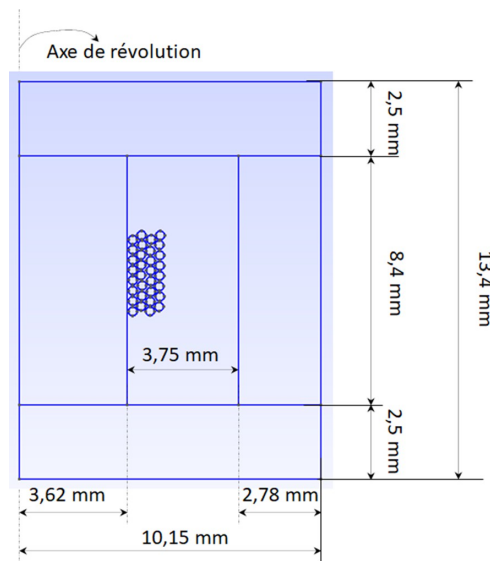


**Figure 4.**  
*Geometry taken into account to illustrate the propagation of uncertainties related to the positions of conductors in the winding window.*

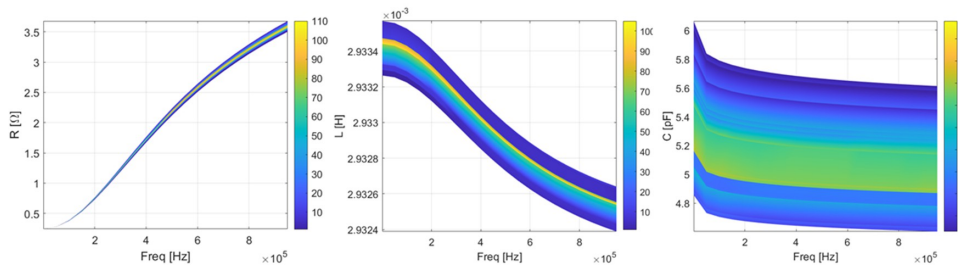


**Figure 5.**  
 Evolution of RLC parameters superimposed on their distributions in the form of colored map, as functions of frequency.

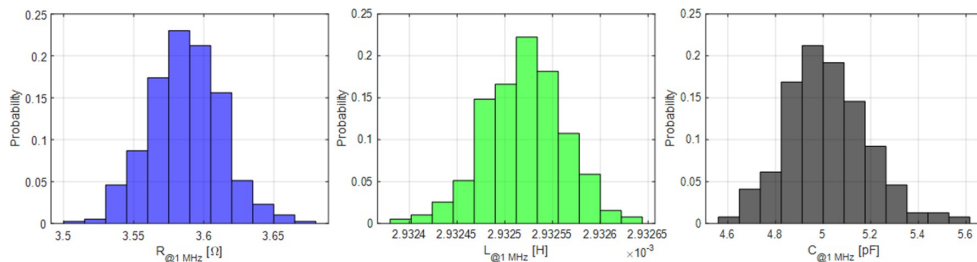


**Figure 6.**  
 RLC parameter histograms (with probability densities on the ordinate) taken at 1 MHz.

of different parameters are superimposed in the form of colored maps to their evolutions. Their scale represents the number of realizations of each value associated with each parameter.

The stopping criterion for this MC simulation is based on the convergence of the mean of each parameter of interest. According to this criterion, one can see from **Figure 6** that the number of MC iterations is sufficient. A zoom at 1 MHz on the distributions of RLC parameters can be observed in **Figure 7**. We can observe in the latter normal distributions, which reflect the nature of uncertainties linked to the positions of conductors in the winding window.

### 3. Case study

In this work, we analyze a typical commercial inductor (MCSCH895-680KU, as illustrated in **Figure 8**). It is made up of a NiZn ferrite core type, with a 44-turn winding distributed over four layers. The first three layers include 12 turns per layer. The conductor diameter is 0.37 mm. According to the manufacturer, this inductor is characterized by a nominal inductance value of  $68 \mu\text{H} \pm 10\%$ . The material properties of the ferrite core are modeled using a *Debye* relaxation model along the frequency range of study [11]. In this model, the random aspect is represented by static quantities for which the manufacturer often provides the relative error (20% for magnetic permeability and 10% for electrical conductivity). The dimension of the random input vector (set of geometrical and material uncertainties) of the studied inductor numerical model has been reduced to seven, among which are the real and

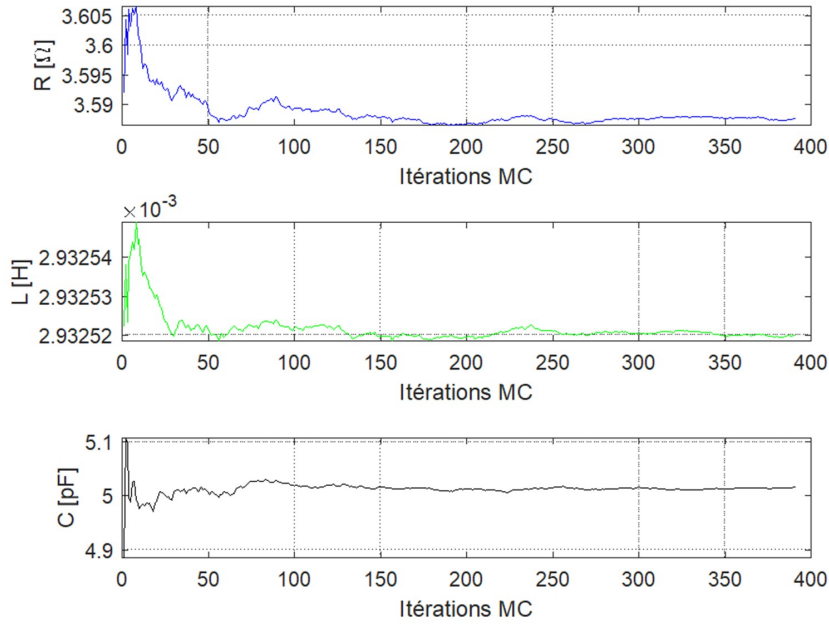


**Figure 7.** Evolution of average RLC parameters measured at 1 MHz as a function of the number of MC iterations (stopping criterion).

**MCSCH895-680KU.**  
INDUCTOR, 68UH, +10%



**Figure 8.** Illustration of the analyzed component.

imaginary parts of  $\underline{v}_{prox}$ ,  $\underline{Z}_{skin}$ , and  $\underline{\mu}_{core}$  and the ferrite core conductivity  $\sigma_{core}$ . Results of the sensitivity analysis will be presented in the section below.

#### 4. Simulation results

To build the stochastic surrogate model based on PCE (using the least angle regression algorithm implemented in UQLab, see Appendix A), 295 calls of the numerical model were sufficient. In order to transform geometric uncertainties into material ones and to extract  $RL$  parameters, the homogenization technique has been implemented in GetDP and Gmsh [12] as computation tools for FE formulations. For experimental validation,  $RL$  parameters were measured across a wide range of

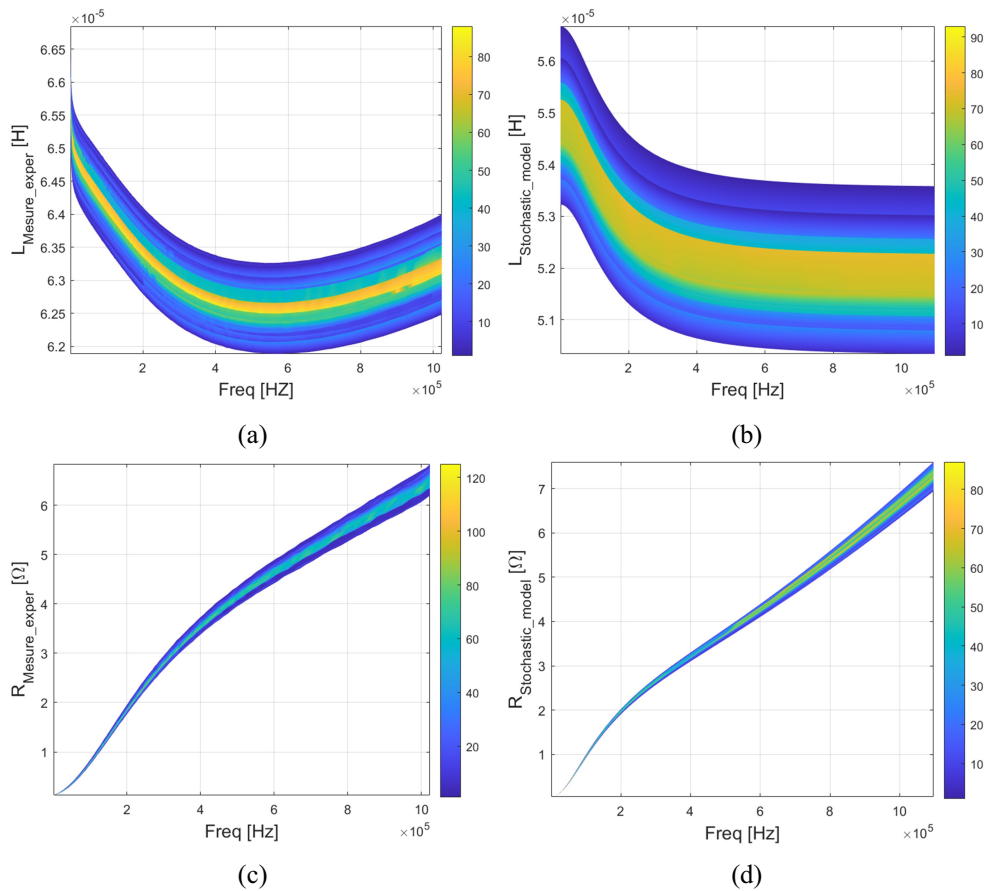


**Figure 9.** Evolution of RL parameters superimposed with their distributions in terms of colored maps, experimentally measured on a batch of 299 MCSCH895-680KU inductors (left) and extracted from stochastic surrogate models (right).

frequencies using a Wayne Kerr impedance analyzer [13]. **Figure 9** shows good agreement between the two evolutions of RL parameters. One can observe an increasing dispersion with frequency in both evolutions (experimental measurement and simulation), especially in the  $R$  parameter, due to the occurrence, with increasing frequency, of skin and proximity effects as well as Joule losses in the ferrite core.

**Figure 10** presents distributions of RL parameters at 537 kHz and 1 MHz. Regarding the  $R$  parameter, distributions overlap with increasing frequency due to the increasing dispersion associated with skin and proximity effects. On the other hand, the increasing dispersion is less marked in the case of the  $L$  parameter. From the order of magnitude of this parameter (less than 0.1 mH) and its low relative error (10% according to data provided by the manufacturer and experimental measurements), we understand why we do not observe any overlap between the distributions associated with the experimental measurements and the stochastic model of this parameter. This parameter exhibits low dispersion over the range of frequencies studied and thus retains a quasi-deterministic

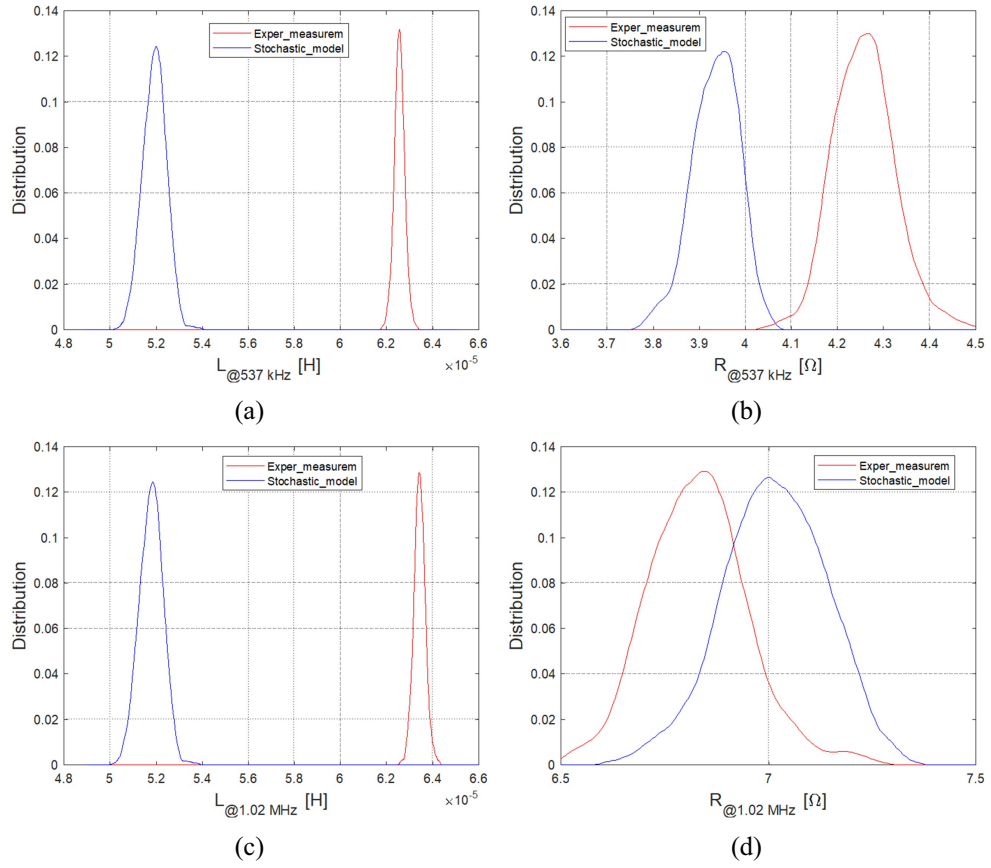


**Figure 10.** Comparison between the distributions of RL parameters from experimental measurements and the stochastic model for frequencies 537 kHz (above) and 1.02 MHz (below).

aspect. From such behavior, one should expect low sensitivity to the uncertainties that are taken into account in the stochastic model (as discussed below).

The post-processing of the obtained PCE allowed for sensitivity analysis to be carried out through the computation of *Sobol* indices. The evolution of *Sobol* indices related to each component of the reduced input random vector over a wide range of frequencies is presented in **Figure 11**.

The  $L$  parameter seems less influenced by variations in random inputs. This may explain its quasi-deterministic behavior. Only three input variables ( $X_3$ ,  $X_4$ , and  $X_5$ ) have an influence on this parameter. The dependence with respect to  $X_5$  is obvious because the latter expresses the stored magnetic energy in the inductor, which decreases with frequency. This decrease may be explained by the circumvention of conductors by field lines when the operating frequency increases. On the other hand, the dependence with respect to  $X_3$  and  $X_4$  clearly demonstrates the significant influence of geometrical uncertainties on the  $L$  parameter.

Regarding the  $R$  parameter, one can observe the influence of several random input variables. This may also justify the strong dispersion of this parameter and the increase in dispersion with frequency. This is linked to the occurrence of several

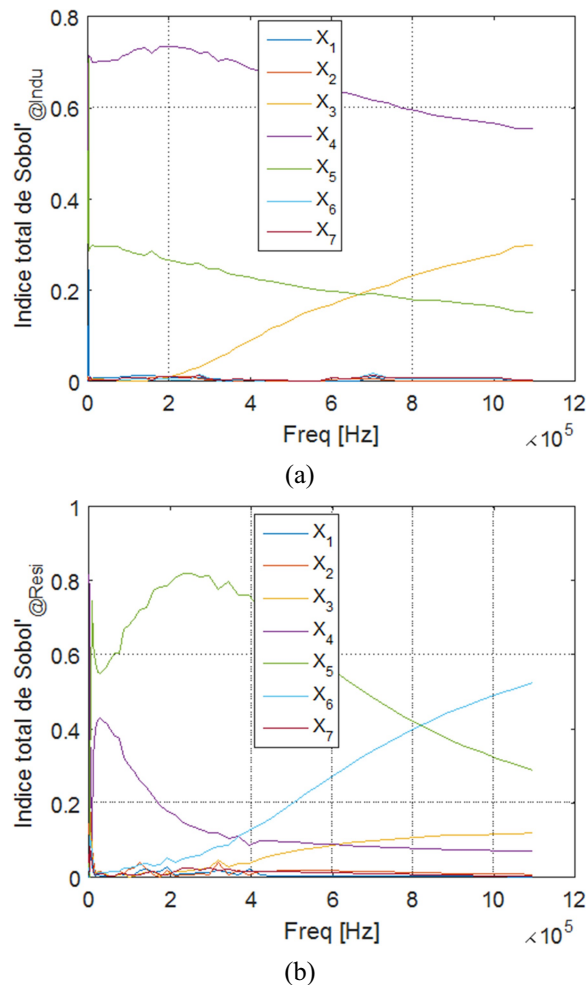


**Figure 11.**  
 Total Sobol indices for stochastic models of  $L$  (above) and  $R$  (below).

phenomena with increasing frequency, including skin and proximity effects through conductors in the winding.

Its dependence on imaginary parts  $X_2$ ,  $X_4$ , and  $X_6$  is obvious since the latter are images of loss (in conductors and the ferrite core). However, its dependence on  $X_7$ , which is almost constant with frequency, is related to eddy current losses in the ferrite core, which normally should decrease with frequency.

One can clearly notice that the  $R$  parameter is strongly influenced by material uncertainties, including the material properties of conductors and the ferrite core.

## 5. Conclusions

In this chapter, we carried out a sensitivity analysis of  $RL$  parameters of a wound inductor according to the considered random inputs in our study. To deal with the

high dimension of the random input vector, we transformed geometrical uncertainties (linked to the positions of conductors in the winding window) into material ones, and we computed a surrogate model based on PCE across the wide range of frequencies in the study. This approach allowed us to compute *Sobol* indices related to the sensitivity analysis of output parameters with respect to the new random input vector, whose dimension has been reduced to seven. Our case study was conducted on a specific wound inductor (MCSCH895-680KU) to obtain reference values for computed output parameters from the numerical model used. Based on the sensitivity analysis, we found that the  $L$  parameter is mostly influenced by geometrical uncertainties (positions of conductors in the winding area), while the  $R$  parameter is mostly influenced by material uncertainties linked to the electrical and magnetic properties of conductors (electrical conductivity) and the ferrite core (magnetic permeability and electrical conductivity). Looking ahead, in the case of a non-symmetrical inductor geometry, it could be challenging to model in 3D the geometric uncertainties associated with the winding and to study their influence on the local electromagnetic behavior via the  $C$  parameter.

## Appendices

### Appendix A: Different strategies for PCE building

The PCE is the development  $\mathcal{M}^{PC}(X)$  of a model response  $Y = \mathcal{M}(X)$  in the space of random functions relating to the distributions of its random inputs  $X$ .

$$Y = \mathcal{M}(X) \approx \mathcal{M}^{PC}(X) = \sum_{\alpha \in \mathcal{A}} y_{\alpha} \varphi_{\alpha}(X), \quad (8)$$

where  $\mathcal{A}$  is the truncated set of multi-indices of the PCE.

The computation of its coefficients  $y_{\alpha}$  is carried out by different methods, including the most classic projection method [3], which uses multiple integrals. The number  $N$  of integration nodes depends on the dimension  $M$  of the random input and on the maximum degree  $p$  of the PCE according to the expression:

$$N = (p + 1)^M \quad (9)$$

Hence, this explosion in the number of evaluations of the numerical reference model is often called the *curse of dimensionality*.

To build the PCE substitute, we start with a sample of size  $N$  of the random inputs (called the experimental design). The determination of the PCE coefficients then involves solving the following optimization problem:

$$\hat{y} = \arg \min \mathbb{E} \left( y^T \varphi(X) - \mathcal{M}(X) \right)^2, \quad (10)$$

For which the solution is given by the ordinary least squares (OLS) method, namely,

$$\hat{y} = (A^T A)^{-1} A^T y, \quad (11)$$

where  $A_{ij} = \varphi(X)_j(x^{(i)})$  contains the values taken by the polynomial basis at the level of points of the experimental design and is called the experimental matrix  $i = 1, \dots, N; j = 0, \dots, P - 1$ .

An important parameter for evaluating the quality of the constructed PCE is the relative empirical error, deduced from the experimental design. This kind of parameter can lead to over-fitting, since it decreases with the degree  $p$  of the PCE regardless of the size  $N$  of the experimental design. In order to overcome this problem, an adaptive

technique for constructing sparse PCEs using an *a priori* cross-validation error was used in this work. This type of PCEs implies low-order interactions between the random variables, which are, by experience, the most important in various numerical models (sparsity-of-effects principle). This was taken into account in the initial optimization problem (see Equation [6]) by associating it with a penalty term of the form  $\lambda \|y\|_1$ , that is, the problem:

$$\hat{y} = \arg \min_{y \in \mathbb{R}^{|\mathcal{A}|}} \mathbb{E}[(y^T \varphi(X) - \mathcal{M}(X))^2] + \lambda \|y\|_1, \quad (12)$$

where the regularization term  $\|y\|_1 = \sum_{\alpha \in \mathcal{A}} |y|_\alpha$  forces the minimization to favor low-rank solutions. In this latter context, the least angle regression (LAR) algorithm is the one implemented in UQLab [14], a specific module for uncertainty quantification under MATLAB.

From **Table A1**, we can observe the accuracy with which we manage to fit the parameter  $R$  (from the FE method) at 1.2 MHz, thanks to the LAR strategy. For the same maximum degree, one observes a sparser basis than with the OLS strategy.

| $R_{@1.2\text{MHz}} [\Omega]$ | Error  | Max. degree | Non-zero coefficient. |
|-------------------------------|--------|-------------|-----------------------|
| OLS                           | 0.4723 | 2           | 15                    |
| LAR                           | 0.0406 | 2           | 8                     |

**Table A1.**  
Comparing OLS and LAR strategies for the surrogate model of  $R_{@1.2\text{MHz}}$  parameter.

## Author details

Geoffrey Lossa<sup>1,2\*</sup> and Olivier Deblecker<sup>2</sup>

1 ISPT-Kinshasa, Kinshasa, DR Congo

2 University of Mons, Mons, Belgium

\*Address all correspondence to: [glossa@isptkin.ac.cd](mailto:glossa@isptkin.ac.cd)

## IntechOpen

© 2025 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Lossa G, Deblecker O, De Grève Z. Influence of geometric uncertainties on the RL parameters of wound inductors modelled using the finite element method. *IEEE Transactions on Magnetics* Miami. 2017;**53**(6)
- [2] Sudret B. Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering and System Safety*. 2008;**93**(7):964–979
- [3] Marelli S, Luthen N, Sudret B. Polynomial chaos expansions - UQLab user manual. In *Chair of Risk, Safety and Uncertainty Quantification*. ETH Zurich, Switzerland; 2022
- [4] Otto F, Gloria A, Duerinckx M. Characterization of fluctuations in stochastic homogenization, *Workshop on Uncertainties Quantification*, 3rd GAMM AGUQ, 2018
- [5] Meunier G, et al. Homogenization for periodical electromagnetic structure: Which formulation? *IEEE Transactions on Magnetics* 2010;**46**:3409–3412
- [6] Vazquez Sabriego R, et al. Time-domain homogenization of windings in 3-D finite element models. *IEEE Transactions on Magnetics* 2008;**44**:3989–3991
- [7] Lossa G, Deblecker O, De Grève Z, Comparing Two Approaches for Generating Realistic Patterns in the 2D Stochastic FE Modeling of Wound Inductors, 2020. Available from: <https://www.researchgate.net/publication/343335404ComparingtwoApproachesforGeneratingRealisticPatternsinthe2DStochasticFEModelingofWoundInductors>
- [8] Lossa G, Deblecker O, Geuzaine C, De Grève Z, Building fast stochastic surrogate models for extraction of RL parameters of wound inductors modelled using FEM, *IEEE Xplore*, CEFC 2021, Pisa
- [9] Wagner P-R, Nagel J, Marelli S, Sudret B. *Bayesian Inference for Model Calibration and Inverse Problems - UQLab User Manual*. ETH Zurich, Switzerland: Chair of Risk, Safety and Uncertainty Quantification; 2022
- [10] Marelli S, Sudret B, UQLab: A framework for uncertainty quantification in Matlab, *Int. Conf. on Vulnerability, Risk analysis and Management*, Proc. 2nd ICVRAM, Liverpool, UK, 2014, 2554–2563
- [11] Lossa G, Deblecker O, De Grève Z, Experimental Validation of Geometrical and Material Uncertainties Consideration in the Modeling of Wound Inductors Using FEM. In: *Photonics & Electromagnetics Research Symposium (PIERS)*; 2021; Hangzhou, China. p. 69–73. DOI: 10.1109/PIERS53385.2021.9694904
- [12] Geuzaine C, Remacle J-F. Gmsh: A three-dimensional 3-D finite element mesh generator with built-in pre- and post-processing facilities. *International Journal for Numerical Methods in Engineering*. 2009;**79**(11):1309–1331
- [13] Lossa G, Deblecker O, De Grève Z. Influence of material uncertainties on the RLC parameters of wound inductors modeled using the Finite element method. *Open Physics*. 2018;**16**(1): 227–231

[14] Marelli S, Sudret B, UQLab user manual – Polynomial chaos expansions, Chair of Risk, Safety and Uncertainty Quantification, Report # UQLab-V1.3-104, ETH Zurich, Switzerland, 2019



*Edited by Mahboub Baccouch*

The finite element method (FEM) is a powerful and versatile computational tool widely used to approximate solutions of partial differential equations in science, engineering, and technology. This book provides a comprehensive introduction to the principles, formulations, and applications of FEM, offering readers a unified approach to modeling complex physical phenomena. From foundational concepts to advanced numerical techniques, it presents practical strategies for solving problems in structural analysis, heat transfer, fluid dynamics, microfluidics, wearable sensor design, and electromagnetic simulations. Emphasizing both theoretical rigor and real-world implementation, the book guides readers through efficient modeling, mesh generation, material and multiphysics coupling, and uncertainty quantification, highlighting best practices for achieving accurate, reliable, and computationally effective results. Designed for students, researchers, and practitioners, it balances conceptual understanding with practical application, making it an essential resource for anyone looking to apply FEM to engineering, scientific, or technological challenges. With clear explanations, illustrative examples, and a focus on modern computational tools, this volume equips readers with the knowledge and techniques needed to tackle complex problems and advance innovation across a wide range of disciplines.

Published in London, UK

© 2026 IntechOpen  
© vsijan / nightcafe.studio

**IntechOpen**

