NPL Report CMSC 29/03

Report to the National Measurement System Policy Unit, Department of Trade and Industry

Model Validation in Continuous Modelling

Trevor Esward, Gabriel Lord, and
Louise Wright

September 2003

NPL REPORT CMSC 29/03

Model Validation in Continuous Modelling

Trevor Esward*, Gabriel Lord†, and Louise Wright*
* Centre for Mathematics and Scientific Computing, National Physical Laboratory
† Department of Mathematics, Heriot-Watt University

September 2003

ABSTRACT

This report addresses issues connected with validation of continuous models for metrology. It provides general advice about the process of validation and describes methods and techniques that can be used for validating continuous models, many of which have been specifically developed to address problems peculiar to continuous modelling. The report also aims to provide simple examples of the use of each technique, and to identify situations where the techniques will not work or are not sufficient on their own. It offers advice on some of the common causes of error identified by some of the methods, and suggests ways of avoiding or removing them.

ISSN 1471-0005

National Physical Laboratory
Queens Road, Teddington, Middlesex, TW11 0LW

Approved on behalf of the Managing Director, NPL
by Dave Rayner, Head of the Centre for Mathematics and Scientific Computing

Contents

1 Introduction

This report addresses issues connected with the validation of continuous models for metrology. In particular, it provides general advice about the process of validation and describes methods and techniques that can be used for validating continuous models. Many of these methods have been specifically developed to address problems peculiar to continuous modelling, although some are more general. The report also aims to provide simple examples of the use of each technique, and to identify situations where the techniques will not work or are not sufficient on their own. It offers advice on some of the common causes of error identified by some of the methods, and suggests ways of avoiding or removing them.

The conclusions of a report on continuous modelling in metrology [28], produced as part of an earlier Software Support for Metrology (SS*f*M) programme, made a number of recommendations of topics within continuous modelling that required further investigation. In particular, model validation and uncertainty estimation were mentioned as being key requirements for the application of continuous modelling techniques to metrological problems. This report takes those conclusions [28] as a starting point, and investigates some of the issues that were raised.

The contents of this report were also motivated by the results of a survey of SS*f*M club members. The survey requested details of their continuous modelling work, including asking if there were areas where they felt more guidance would be useful. Many respondents felt that guidance on validation and error estimation would be useful, and so this report provides practical guidance and examples of application of validation methods.

This report is strongly linked to the companion reports “Uncertainty Evaluation In Continuous Modelling” [24] (as ideally a final validation should include an estimate of the uncertainty of the results of the model) and “Guide to the use of finite element and finite difference software” [25] (as some of the errors highlighted during validation can be avoided by the correct use of software packages). It also has strong links with activity in the SS*f*M project “Testing Continuous Modelling Software” [26], since some of the techniques listed here check for errors caused by incorrect software rather than incorrect modelling assumptions.

The final section of this report contains a summary and review of the material presented. As the report is quite lengthy, it is recommended that readers begin with the summary in order to identify the sections of the report that are of most interest and relevance to their work.

1.1 What is validation?

In the Best Practice Guide to Discrete Model Validation [1], produced as part of the first SS*f*M program, model validation was defined as

“Determining whether the results produced are consistent with the input data, theoretical results, reference data, etc. All stages need to be examined. Does the model adequately encapsulate what is known about the system? Does the method of solution produce unbiased estimates of the parameters and valid estimates of uncertainties?”

These ideas are largely applicable to the validation of continuous models, but some points need further investigation. The main purpose of discrete models as examined in the Guide was to determine parameters and their appropriate levels of uncertainty from experimental data. In the survey of SS*f*M Club members, respondents were asked how often they used continuous modelling for a variety of tasks such as experimental design and data processing. The most frequent use of continuous modelling was to gain a deeper understanding of the process being modelled. This means that the determination of parameters and uncertainties is not the only issue of concern, and that consideration needs to be given to other phenomena where possible. The consideration of other phenomena is partly what has motivated the investigation of “looks right” as a validation criterion, since the behaviour of many phenomena of interest is known in approximate terms but not detailed ones. For example, often experimental experience leads to knowledge of practical limits on some variable which can help in determining the correctness of a model.

In the Guide [1], discrete model validation was considered as having two aspects:

- **internal consistency checks**, which validate the mathematical correctness of the model description. A model is considered to be internally consistent if its outputs are valid so long as its inputs are valid. Checking internal consistency involves checking for the mathematical correctness of the modelling process, for instance checking that linearization is reasonable for the operating ranges of the model or checking that any other approximations used are reasonable.
- **external consistency checks**, which validate the model results against other sources of information such as experimental results or other prior information. A model is considered to be externally consistent if its outputs are not contradicted by any other valid information. Checking external consistency involves ensuring that the model represents the real situation, by considering whether the assumptions made are valid for this particular system, and by comparing model results with experimental data or other information.

Many of these concepts are useful in continuous model validation, but continuous modelling could be considered as needing an extra consistency check. Discrete modelling usually consists of developing suitable equations and then solving them or finding best fit parameters from them. Continuous modelling usually consists of identifying suitable continuous differential equations, developing a discrete form of the equations so that they can be solved, and then solving them. The discretisation step of this process means that a key part of continuous model validation is checking the consistency of the discrete form with the continuous form, including checking the adequacy of the mesh associated with the discretisation. The importance of this aspect leads to the definition of **inter-model consistency checks** as checks that the chosen discretisation method adequately describes the original continuous model. A continuous model and its discrete approximation have inter-model consistency if the solution to the discrete approximation converges to the solution of the continuous model as the step size of the discretisation tends to zero. Such checks are common practice in the development of finite approximations to continuous problems.

Some of the information in the Guide [1] can be useful for validation of the inputs to continuous models. In particular, if the model is to be used for parameter determination and an uncertainty is required then the section on statistical model validation will be useful to check the validity of the assumptions about the input parameters of the model. This report does not examine statistical model validation, but issues connected with statistical continuous models and their validation are addressed in the report Uncertainty Evaluation in Continuous Modelling, a deliverable of Work Package 3 of this project.

Parameter estimation

As has been mentioned above, the use of continuous models for parameter estimation is comparatively rare (although it does occur: an example of its use in material property estimation is available [22]). This rarity is probably due to the combination of lengthy run times and discretised equations making the choice of optimisation methods very limited. Due to this rarity, the issue of estimator validation is not addressed specifically in this report. However, the recommendations for discrete model estimator validation in section 3 of the Guide [1] are directly applicable to continuous models as well. This application of techniques used in discrete model validation is possible partly because most of what is said does not rely on any particular properties of discrete models, and partly because a discretised version of a continuous model can be regarded as a discrete model for many purposes. Users of continuous modelling for parameter estimation are recommended to consult the Guide for this information.

The Structural Dynamics 2000 Forum contained a special session examining issues around validation of finite element models, and the keynote address of that session [2] made a number of points that are relevant to the material in this report:

- "It must somehow be verified that the many assumptions involved in the successive steps of idealization, discretisation and modelling yield satisfactory predictions"
- "Reproducing test data does not guarantee predictability away from the region in the design space that relates to the test data"
- "It is our opinion that the focus of research in model validation should be shifted from validating deterministic models to validating statistically accurate models… Therefore, the concept of validation should be strongly coupled to uncertainty quantification…"

The first of these points is equivalent to the definitions of validation quoted above [1], and includes the point that the discretisation needs to be checked. The second point is particularly important to consider when validating models that are to be used for prediction using inputs that have not been used in experiments. It does not mean that models cannot be used for predictions away from the areas in which they have been validated, but it does mean that use of the model in such cases needs to be considered carefully. In particular, it is important to check that none of the assumptions made during model development have been violated in the new region of the design space.

Uncertainty quantification

The final point is very important for metrology applications, where the concept of uncertainty quantification is already well-established. Input parameters of models are generally derived from measurement, and so they have associated uncertainties. Thus, model results based on these input parameters will also have uncertainties. Any experimental data used for validation will have associated uncertainties, and the ideal validation process would take these uncertainties into account to produce a probabilistic indication of the extent to which a model is validated.

These concepts have already been accepted in discrete modelling for metrology, but they are considerably less commonly considered in continuous modelling. In discrete modelling, the uncertainties of the input parameters lead to the definition of their joint probability density function (pdf). This joint pdf is then propagated through the equation linking the input parameters with the model results to obtain an estimated pdf for the results of interest. Usually this propagation is carried out either analytically, or using appropriate approximation methods, or by using Monte Carlo simulation.

There are several contributing factors to the lack of use of uncertainties in continuous modelling. One is that most continuous models are quite large in terms of the number of variables and are often non-linear and so propagation of distributions using analytic methods is usually prohibitive in terms of complexity. Similarly, they often take a long time to run so methods such as Monte Carlo simulation are prohibitive in terms of time. Another factor is that many models are developed using proprietary "black-box" software, and the use of such software means that the modeller does not have access to the derivatives of the functions or expressions comprising the model that would help with the determination of the sensitivity coefficient typically used in uncertainty evaluation [23]. Additionally, black box software does not encourage the consideration of uncertainty in its inputs, and often user-friendly packages with high quality graphics can encourage users to think of their results as a unique and correct answer rather than part of a range of possible answers depending on the distributions of the input parameters.

Validation is a part of model development and improvement, and so is not just about passing or failing consistency checks. The results of some validation methods provide a quantitative estimate of the error generated by an assumption or approximation, and these estimates can be used to indicate ways of improving the model, for instance by using a finer mesh or by implementing a more detailed model of some phenomenon. Some of the external consistency checks can lead to new parameter values being suggested as improvements to the model, although care should be taken to ensure that the uncertainty of the measurement data is taken into account in such cases.

In the light of these points, continuous model validation could be defined as determining whether the results produced are consistent with the input data, theoretical results, and reference data, including checking of

- adequacy and mathematical correctness of the model description (internal consistency)
- adequacy of the discretisation chosen in terms of convergence and estimation of discretisation errors (inter-model consistency)
- correctness of results against other sources of information such as experiment or other prior information (external consistency)

Wherever possible, this process should generate uncertainty and error estimates as well as an overall pass or fail response for the model.

1.2 Structure of the report

This report consists of ten main sections, four of which contain technical details of methods and examples of their application. The other sections put the methods into context, explain the potential pitfalls of their use, and explain some of the concepts needed to understand the methods fully.

To make the worked examples of applications for the methods as clear as possible, the same notation has been used throughout. Section 2 explains this notation and introduces various concepts and definitions that are used in the explanation of some of the methods. Some of the material in this section will be familiar to readers with a strong mathematical background.

Section 3 describes the main sources of error that occur in continuous modelling. Some of these sources of error are not discussed further in this report, so the section explains why these sources have been neglected and mentions other sources of information about these types of error.

Section 4 contains general advice on model validation. This advice includes consideration of multiple solutions and non-uniqueness and advice on planning model validation to ensure all of the important aspects of the model are checked.

As was mentioned in section 1.1, the methods used for validating continuous models can be split into three classes: internal consistency checks, external consistency checks, and inter-model consistency checks. Sections 5, 6, and 7 of the report describe some of the methods within each class, with section 5 covering internal checks, section 6 external checks, and section 7 inter-model checks. Section 8 gives a worked example of the full model validation process using a real metrological problem as the example. This example uses as many of the methods described in sections 5, 6, and 7 as possible to validate the model fully.

Section 9 summarises the key points mentioned in the report, and section 10 gives the full list of references mentioned in the text.

2 Notation and nomenclature

Throughout this report, the same notation is used for vectors, functions, and discretised versions of continuous problems. The notation used is common in this branch of mathematical modelling, but is explained fully in this chapter to avoid confusion. This chapter also includes background information about norms and inner products and defines some of the terms used in the sections on a priori methods. It can probably be skipped by anyone with a mathematical background.

Additionally it should be noted that, unless otherwise stated, graphs of results have been created from results at discrete points joined by straight lines, and no attempt has been made to display the behaviour between the points. Lines are generally included to make the graphs clearer rather than to represent the solution at intermediate points.

Some of the proofs and theorems given here require conditions on the smoothness of the solution. In some cases it is clear what the requirements are. For instance, any proof that involves expansion of the solution in terms of a Taylor series requires existence of all partial derivatives appearing in that series. In general, if the proof requires conditions on the solutions that are any stronger than a reasonably smooth solution, they will be stated. Most metrological problems have reasonably smooth solutions once any noise has been removed, so these conditions are likely to hold for most real applications.

2.1 Notation

Throughout the following report, unless otherwise stated, $\mathbf{x}$ is a column vector, $\mathbf{x}^{\mathrm{T}}$ is its transpose, a row vector, and $\mathbf{x}^{\mathrm{T}} = \{x_1, x_2, \ldots x_n\}$. There are circumstances under which $\mathbf{A}$ can indicate a matrix, and if this is the case it will be stated. R^N is real N-dimensional space and C^N is the complex equivalent.

For a function $u(\mathbf{x}, t) = u(x, y, z, t)$, the discretised approximation is written

$$U_{i,j,k}^{n} \approx u(i\Delta x, j\Delta y, k\Delta z, n\Delta t)$$

where the mesh is assumed to be uniformly spaced in all dimensions including time. The superscript always denotes the time index, so if the problem is steady-state the superscript is neglected. In some cases (such as finite element problems) the mesh points may be numbered individually, in which case $U^n{}_i \approx u(\mathbf{x}_i, n\Delta t)$. Similarly, if the time steps are not uniform then $U^n{}_i \approx u(\mathbf{x}_i,\ t_n)$. The vector of the approximate solution at all mesh points at time t_n is written $\mathbf{U}^n$.

The region of space within which the problem is to be solved, called the **domain**, is denoted Ω and is a subset of R^n, its boundary is $\partial\Omega$, and the time interval is usually taken to be finite and is written $[0, t_F)$, so the space-time domain is written $\Omega \times [0, t_F)$. In most cases R^n will be $n = 2$ or 3.

For stress problems, σ_{xx} is a normal stress in the x direction, τ_{xy} is a shear stress, and the displacements are written as $\mathbf{u}^{\mathrm{T}} = \{u, v, w\}$.

2.2 Consistency, stability, and convergence

Consistency, stability and convergence are key terms in the numerical analysis of discrete approximations to continuous differential equations, as they define the properties that any useful method must have [3, Chapter 5].

Suppose that $u(\mathbf{x}, t)$ is the exact solution to the problem

$$C\frac{\partial u}{\partial t} = L(u) \text{ on } \Omega\times(0,t_F], \qquad (1a)$$

$$g(u) = g_0 \text{ on } \partial\Omega_0 \subset \Omega, \qquad (1b)$$

$$u = u^0(\mathbf{x}) \text{ on } \Omega \text{ at } t = 0, \qquad (1c)$$

where C is a constant that could be zero (in which case the initial conditions (1c) are not required) and L and g are linear differential operators. See section 2.3 for a definition of a linear differential operator.

Suppose that an approximate discretised solution is generated on an evenly-spaced mesh and $\mathbf{U}^n$ let be the approximate solution at time $n\Delta t$, as explained in section 2.1. The $\mathbf{U}^n$ are generated by some linear method so that

$$\mathbf{AU}^{n+1} = \mathbf{BU}^n + \mathbf{F}^n \ , \qquad (2)$$

where **A** and **B** are matrices and $\mathbf{F}^n$ is a vector, $\mathbf{U}^0$ is given from the initial conditions (1c)**,** and **A**, **B**, and the $\mathbf{F}^n$ depend on Δx, Δy, Δz, and Δt. If the problem (1) is not time-dependent then the formulation will still be valid with $\mathbf{A} = \mathbf{0}$**.** Some property h of the mesh, that describes a useful measure of size, can be defined. Examples are the maximum value of Δx, Δy, and Δz, the minimum diameter of a circle surrounding all mesh points related to some chosen point, or the maximum of Δx, Δy, and Δz after rescaling to take varying typical time derivatives into account.

The **truncation error** $\mathbf{T}^n$ of a method is defined as the error obtained on substituting the exact solution into the method, i.e. $\mathbf{T}^n = \mathbf{Au}^{n+1} - (\mathbf{Bu}^n + \mathbf{F}^n)$, where $\mathbf{u}^n$ is the exact solution evaluated at the mesh points ordered in the same way as $\mathbf{U}^n$. This truncation error is effectively the same as the spatial discretisation error mentioned in section 3.2. The method (2) is **consistent** with problem (1) if

$$T^n_{i,j,k} \to 0 \ \forall\, i,j,k \text{ as } h \to 0 \text{ and } \Delta t \to 0.$$

Stability here is a property of time-dependent problems only. If the method (2) is applied to two different sets of initial conditions, $\mathbf{V}^0$ and $\mathbf{W}^0$, then it is stable if there is a constant K that is independent of h, Δt, $\mathbf{V}^0$ and $\mathbf{W}^0$ such that

$$\left\|\mathbf{V}^n - \mathbf{W}^n\right\| \le K\left\|\mathbf{V}^0 - \mathbf{W}^0\right\|,\ \forall\, n\Delta t \le t_F,$$

for some norm $\|\bullet\|$ and for t_F as defined in (1a). This definition means that a small change in the initial conditions will not lead to an unbounded change in the solution produced by the method within the time interval of interest. This is linked to the question of **well-posedness** of the problem (1): a well-posed problem is one such that i) a solution exists for any initial data $\mathbf{u}^0$, where $\|\mathbf{u}^0\|$ is bounded, and ii) any two solutions of which obey $\|\mathbf{v} - \mathbf{w}\| \le K' \|\mathbf{v}^0 - \mathbf{w}^0\|$ for all $t \le t_F$.

The method (2) provides a **convergent approximation** to the problem (1) if

$$\left\|\mathbf{U}^n - \mathbf{u}^n\right\| \to 0 \text{ as } \Delta t \to 0, h \to 0, n\Delta t \to t \in (0,t_F]$$

for all initial data for which (1) is well-posed. It should be noted that convergence is only proved over a finite time interval: in general, the error bound will contain a term of the form e^{Kt} which can grow without bound for large times, and so methods proving convergence are not necessarily sufficient for the validation of simulations over a long time period.

2.3 Vector spaces, function spaces, and operators

A **vector space** V over R^n or C^n is a set of vectors such that if $\mathbf{u}$, $\mathbf{v}$ are in V and λ, μ are in R or C as appropriate then $\lambda\mathbf{u} + \mu\mathbf{v}$ is in V. In addition to this rule, there are a number of conditions on the arithmetic operations in the space. Addition must be associative and commutative, scalar multiplication must be associative, and scalar and vector sums must be distributive. Finally, there must be a zero member $\mathbf{0}$, additive inverses $\mathbf{-u}$, and a multiplicative identity 1, such that $\mathbf{u} + \mathbf{0} = \mathbf{0} + \mathbf{u} = \mathbf{u}$, $\mathbf{u} + -\mathbf{u} = \mathbf{0}$, and $1\mathbf{u} = \mathbf{u}$. The real numbers are a vector space, as are R^n and C^n, but the definitions allow for more unusual sets as well.

A **norm** $\|\bullet\|$ on a vector space V is a map $\|\bullet\|$: $V \to R$ that has three qualities:

- $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x}$ in V, and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} \equiv \mathbf{0}$,
- $\|k\mathbf{x}\| = |k|\|\mathbf{x}\|$ for all scalars k,
- $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for all $\mathbf{x}$,$\mathbf{y}$ in V.

Some examples of norms are given in section 2.4. Later in this section it will be shown that other useful norms can be defined from positive definite symmetric bilinear maps.

A **linear map** F between two vector spaces V and W is such that for all vectors $\mathbf{u}$, $\mathbf{v}$ in V and all scalars λ, μ, $F(\lambda\mathbf{u} + \mu\mathbf{v}) = \lambda F(\mathbf{u}) + \mu F(\mathbf{v})$ is in W. A **bilinear map** B:$V \times V \to W$ is such that $B(\mathbf{u},\mathbf{v})$ is linear in $\mathbf{u}$ for fixed $\mathbf{v}$ and linear in $\mathbf{v}$ for fixed $\mathbf{u}$. A **symmetric** bilinear map has $B(\mathbf{u},\mathbf{v}) = B(\mathbf{v}, \mathbf{u})$ for all $\mathbf{v}$, $\mathbf{u}$ in V, and a **positive definite** bilinear map has $B(\mathbf{u},\mathbf{u}) > 0$ for all $\mathbf{u} \neq \mathbf{0}$.

Consider a real positive definite symmetric bilinear map B:$V \times V \to R$. Since B is positive definite, $\surd B(\mathbf{v}, \mathbf{v}) > 0$ for all $\mathbf{v} \neq \mathbf{0}$, and if $\mathbf{v} = \mathbf{0}$ then $\surd B(\mathbf{v}, \mathbf{v}) = 0$ by linearity. This is the first condition required for $\surd B(\mathbf{v}, \mathbf{v})$ to be a norm on V. For any real scalar constant k, $\surd B(k\mathbf{v}, k\mathbf{v}) = \surd\{kB(\mathbf{v}, k\mathbf{v})\} = \surd\{k^2 B(\mathbf{v}, \mathbf{v})\} = |k|\surd B(\mathbf{v}, \mathbf{v})$, which is the second condition necessary for $\surd B(\mathbf{v}, \mathbf{v})$ to be a norm on V. Finally,

$$\begin{aligned}\sqrt{B(\mathbf{u}+\mathbf{v},\mathbf{u}+\mathbf{v})} &= \sqrt{B(\mathbf{u},\mathbf{u})+B(\mathbf{v},\mathbf{v})+2B(\mathbf{u},\mathbf{v})}\\ &\leq \sqrt{B(\mathbf{u},\mathbf{u})}+\sqrt{B(\mathbf{v},\mathbf{v})}+\sqrt{2B(\mathbf{u},\mathbf{v})}\\ &\leq \sqrt{B(\mathbf{u},\mathbf{u})}+\sqrt{B(\mathbf{v},\mathbf{v})},\end{aligned}$$

so that $\surd B(\mathbf{v}, \mathbf{v})$ satisfies all three conditions and can be considered a norm on the vector space V.

Consider the set of functions $F = \{f : \Omega \times [0, t_F) \to R^n$ or $C^n\}$. For $\mathbf{x}$ in Ω and $0 \leq t \leq t_F$, λ a scalar, and f and g members of F, define $(f + g)(\mathbf{x}, t) = f(\mathbf{x}, t) + g(\mathbf{x}, t)$ which will be in R^n or C^n, and $(\lambda f)(\mathbf{x}, t) = \lambda f(\mathbf{x}, t)$. Under these rules, F is a vector space as defined above. A vector space whose vectors are functions is called a **function space**.

Linear maps and bilinear forms can be defined on function spaces in the same way as they are defined for other vector spaces. Linear maps on functions are usually called **operators**. One of the most common types of linear operator in continuous mathematical modelling is the **linear differential operator**. Linear differential operators are used to describe partial and ordinary differential equations, and in R^n they are of the form

$$L\{u(\mathbf{x},t)\}=\sum_{i=0}^{m}\sum_{k_1+k_2+...+k_n+k_{n+1}=i} c_{k_1k_2...k_nk_{n+1}}\frac{\partial^i u}{\partial x_1^{k_1}\partial x_2^{k_2}....\partial x_n^{k_n}\partial t^{k_{n+1}}},$$

where $c_{k_1k_2...k_nk_{n+1}}$ can be functions of **x** and *t*. A simple example is

$$L\{u\}=\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2},$$

but more complicated forms exist. Often, a linear operator *L* acting on a function *u* is written *Lu* without any brackets. Generally, any linear differential equation can be written in the form *Lu* = *f* in some domain $\Omega \times [0, t_F)$.

Consider a linear map L: $F \to F^*$ and a bilinear form B: $F \times F^* \to R$. The **adjoint** of *L* is defined as L^*: $F^* \to F$ such that $B(Lf, g) = B(f, L^*g)$ for all *f* in *F* and all *g* in F^*. *L* is **self-adjoint** if $F = F^*$ and $L = L^*$. For instance, if

$$F = \{u(x)\text{: } R\to R \text{ such that } u\to 0 \text{ as } x\to\pm\infty\},\ L\{u\}=\frac{d^2u}{dx^2},\text{ and } B(u,v)=\int_{-\infty}^{\infty}uvdx,$$

then

$$B(Lu,v)=\int_{-\infty}^{\infty}Luvdx=\int_{-\infty}^{\infty}\frac{d^2u}{dx^2}vdx=\left[\frac{du}{dx}v\right]_{-\infty}^{\infty}-\int_{-\infty}^{\infty}\frac{du}{dx}\frac{dv}{dx}dx$$
$$=0-\left\{\left[u\frac{dv}{dx}\right]_{-\infty}^{\infty}-\int_{-\infty}^{\infty}u\frac{d^2v}{dx^2}dx\right\}=\int_{-\infty}^{\infty}u\frac{d^2v}{dx^2}dx=B(u,Lv)$$

and so *L* is self-adjoint for this function space and bilinear form. If $F = F^*$ and $B(Lf,f) \geq 0$ for all *f* in *F* then *L* is positive definite for that vector spaces and bilinear form. By following a similar argument to that outlined above for real positive definite symmetric bilinear maps, it can be shown that a linear operator *L* that is self-adjoint and positive definite for some bilinear form *B* and function space *F* can define a norm for the functions in *F* by writing $\|u\|_B = \sqrt{B(Lu, u)}$.

2.4 Norms

Many of the techniques listed in the next sections involve comparison between two sets of numbers. In order to tell whether the numbers being compared are sufficiently close for the model to have passed the validation test, it is necessary to have a numerical value for their closeness. This value is analogous to the modulus to express the size of a complex number and the usual expression for the length of a vector. The quantity used is a more flexible definition of size, called a norm, and is usually written $\|\mathbf{x}\|$. The full mathematical definition for a norm is given in section 2.3.

If only two real numbers are being compared, the natural choice for the norm is the absolute value of the difference between them, but there are many different choices possible when the comparison is between two data sets. The most useful choice of norm depends on what aspect of the model is of the most interest to the user.

Suppose throughout the following that we wish to compare two sets of data, **x** and **y**, and write **u** = **x** − **y** for the difference between them. Most norms for vectors with discrete values are of the form

$$\|\mathbf{u}\|_p = \left[\sum_{i=1}^{n} |u_i|^p \right]^{1/p}$$

for some specific value of p. This is called the discrete l_p norm, and is the equivalent of the continuous L_p norm

$$\|f\|_p = \left[\int_{\Omega} |f(\mathbf{x})|^p \right]^{1/p}$$

used for function spaces. Often the sum inside the square brackets is multiplied by $1/n$ to give an average value of the difference.

Weighted sums of the form

$$\|\mathbf{u}\|_p = \left[\sum_{i=1}^{n} |w_i u_i|^p \right]^{1/p}$$

for some set of weights w_i are useful in some cases. Weighted norms are particularly useful when some parts of the results are more important than others, which can occur in optimisation for parameter determination, or where it is desirable to reduce the effects of some results such as those around a singularity.

There are three norms in common usage for metrology problems: the l_1 norm, the l_2 norm, and the l_∞ norm. The l_1 norm is the sum of the absolute values of the differences, the l_2 norm is the root-sum-of-squares norm commonly used for calculating vector sizes, and the l_∞ norm is given by $\|\mathbf{u}\|_\infty = \max\{ |u_i| : i = 1, \ldots n\}$

In general the l_2 norm is probably the most commonly used in metrological applications. This widespread use is partly because this norm has a range of applications in curve fitting and regression where parameters are often determined as being best fits in a least squares sense and so it is a fairly well-understood measure of goodness of fit.

As well as these norms, there are other examples that are of interest in specific areas. In particular, the error energy norm is commonly used for problems in finite element analysis. The norm $\|\bullet\|_B$ discussed in section 2.3 is called the energy norm, because in many physical cases it provides a measure of potential energy in the system being modelled. Additionally, it has a useful application to finite element approximations. If an approximation U to the solution of the continuous problem $Lu = f$ in a domain Ω is sought within some function space F, then the finite element approximation will minimise $B(u\text{-}U, Lu\text{-}LU)$ where

$$B(u,v) = \int_{\Omega} uv d\Omega .$$

This minimal property leads to various useful properties of the error bound in this norm.

3 Sources of error

As was mentioned in section 1, part of the aim of the model validation process is to identify, and if possible quantify, sources of error. These errors can be due to i) an imperfect description of the physical situation, ii) the necessity of solving a simplified version of the mathematical description chosen, or iii) the limited precision of computation. The errors can be considered as falling into five main categories:

- Modelling error: the inadequacy of the chosen continuous equation (generally either a differential or integral equation) and boundary or initial conditions in describing reality
- Space discretisation error: error generated by solving a continuous problem on a discrete mesh
- Time discretisation error: error generated (and accumulated) by solving a continuous problem at discrete time steps
- Parameter errors: errors generated by poorly chosen input parameters
- Linear algebra errors: errors generated during the solution of the discretised system.

These categories will now be considered individually.

3.1 Modelling error

Modelling error is the most difficult error to quantify and reduce, since generally the model being used is the modeller’s best judgement of the physical situation. However, sometimes the model has been simplified deliberately in order to decrease the run time, and such a simplification will lead to a modelling error. Common simplifications include the assumption of symmetry, of constancy of physical properties, and of linearity of boundary conditions. It may be possible to obtain a rough estimate of the modelling error produced by these simplifications by starting with the most complex model and comparing its results with those of a simplified version. It may be possible to estimate which results will be most affected by the simplifications and take this increased error into consideration when drawing any conclusions from them.

Generally, it should be possible to make an estimate of the effects of any simplifications to within an order of magnitude. This can be done by considering the size of the term that the simplification neglects relative to the remaining terms. For example, a model of the stress distribution in a loaded beam may assume that the material is perfectly elastic, whereas in reality some of the deformation will be plastic. If the plastic behaviour is understood but has been ignored to decrease the model’s run time, an estimate of the neglected plastic strains can be made and compared with the calculated strains to check that it is insignificant. It is probably more important to perform a comprehensive qualitative assessment of the differences between reality and the idealised mathematical form than to attempt a quantitative assessment of the few factors that seem to be more tractable.

3.2 Space discretisation error

Space discretisation errors are caused by attempting to approximate a spatial continuum with a set of discrete points and so converting a differential or integral equation into a system of algebraic equations. Generally, a larger number of points used in the approximation will produce a more accurate approximation to the continuous equations, but will also produce a longer run time, will need more computer memory to handle the matrix equations and the solution, and may produce a loss of accuracy due to the effect of the errors caused by the increased numerical computation. This means that there is much to be gained by using a carefully chosen set of points.

Various methods exist for identifying and estimating space discretisation errors, from the simple "halve the mesh size and re-run" test to more sophisticated a priori and a posteriori error estimation methods. Some software packages have automatic mesh evolution tools that refine meshes based on error estimates produced by such methods as the calculation progresses so as to strike a balance between discretisation error and run time.

The transition from a continuum to a discrete mesh also makes it necessary to apply the boundary conditions at discrete points, and this can cause problems. Boundaries with discontinuous conditions make it difficult to treat points of intersection where more than one condition could be applied. Curved boundaries can be difficult to deal with if a finite difference method has been used. It is not immediately obvious how to apply forces at individual mesh points so as to produce a uniform pressure on the face of a finite element. All of these features can be potential sources of mistakes as well as being sources of numerical error. Guidance on how to avoid mistakes in discretizing boundary conditions is given in "Guide to the use of finite element and finite difference software" [25], which is a guide to setting up, running, and interpreting the results of a continuous model using proprietary software packages.

3.3 Time discretisation error

Time discretisation errors are similar to space discretisation errors as they are caused by approximating a continuum with a series of discrete points. However, they can be more problematic because they are accumulated at each time step and can grow exponentially. This means that small changes in the time discretisation can lead to extremely large changes in the computed solution.

For example, consider the one-dimensional heat equation,

$$\frac{\partial u}{\partial t}=\frac{\lambda}{\rho c_p}\frac{\partial^2 u}{\partial x^2} \tag{3}$$

$$0 \le x \le 1, \qquad u(0,t)=0, \quad u(1,t)=0,$$
$$u(x,0)=\begin{cases} 2x, & 0 \le x \le \tfrac{1}{2}, \\ 2(1-x), & \tfrac{1}{2} \le x \le 1. \end{cases} \tag{4}$$

The analytic solution to this problem can be written as $u(x,t)=\sum_{n=1}^{\infty} A_n e^{-\alpha n^2 \pi^2 t}\sin(n\pi x)$, where $\alpha = \lambda/(\rho c_p)$ and A_n are the coefficients in the Fourier sine series of $u(x,0)$.

Alternatively, the equation can be discretised using the most straightforward explicit difference scheme, so that

$$U_j^{n+1}=U_j^n+\frac{\lambda}{\rho c_p}\frac{\Delta t}{\Delta x^2}\left(U_{j+1}^n-2U_j^n+U_{j-1}^n\right), \tag{5}$$

$$U_0^n = 0, \quad U_M^n = 0,$$

$$U_j^0 = \begin{cases} 2j\Delta x, & 0 \le j\Delta x \le \tfrac{1}{2}, \\ 2(1 - j\Delta x), & \tfrac{1}{2} \le j\Delta x \le 1. \end{cases}$$

Linear stability analysis shows that for a given Δx, time steps satisfying $2\lambda\Delta t \le \rho c_p \Delta x^2$ will be stable, but time steps with $2\lambda\Delta t > \rho c_p \Delta x^2$ will not (see section 7.1.2 for further details). Figure 1 illustrates this behaviour.

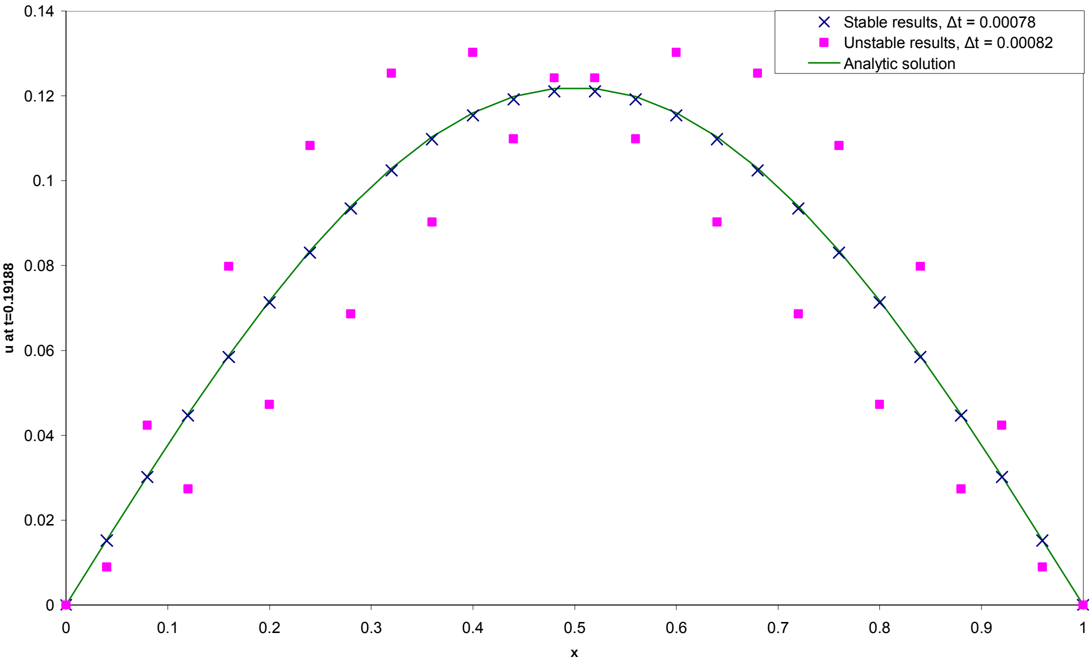

Figure 1: Comparison of analytic solution with results produced using a stable (crosses) and unstable (squares) time step in a simple explicit transient finite difference scheme.

The results were generated using $\Delta x = 0.04$ and $\lambda/(\rho c_p) = 1$ (so that the maximum stable time step is 8.0×10^{-4}) and time steps of 7.8×10^{-4} and 8.2×10^{-4} were used for the stable and unstable results respectively. The results shown are at $t = 0.19188$. It is clear which results are wrong, and the errors grow exponentially as time progresses. This oscillation of the calculated solution about the correct solution is typical of errors caused by an unstable time step or some other form of numerical instability.

Generally, dynamic finite element software have an automatic time step calculator that estimates a stable time step for each element in the mesh and then uses the minimum of all those time steps for the full calculation. The calculation of a stable time step involves a form of a posteriori error estimation for the time discretisation error, using the error estimate for one step to calculate a suitable time step for the next. Some examples are described in sections 7.1.2 and 7.2.2.

Many systems reach a stable state after a long period of time. For instance if no heat source is present, dynamic thermal problems usually tend towards a uniform temperature state, and many dynamic systems reach steady periodic states in the long term. These states are a feature of the physical system and usually of the differential equations used to describe it. Some calculation methods can lead to spurious steady-state solutions that are a feature of the discretised method but not of the original equations, leading to errors in time-dependent predictions. These errors are discussed further in section 4.2.

3.4 Parameter error

Parameter errors are better regarded as parameter uncertainties, since the parameter values used in a model will usually be the best estimate available of the value in reality. As was mentioned in the introduction, there are strong links between model validation and uncertainty quantification. Uncertainty estimation in continuous models is covered in more depth in a companion report, “Uncertainty Evaluation in Continuous Modelling” [23], which includes a report on uncertainty evaluation techniques and several case studies applying them to real metrological problems. In general, methods relate the uncertainty of an input parameter to the uncertainty of some output quantities of interest. The main classes of method considered in the report are sampling techniques and stochastic techniques.

Sampling techniques rely on taking a number of samples from the distributions corresponding to each of the uncertain input values, running the model using these sampled sets of parameters, and using the results to accumulate a distribution for the output quantities of interest. Various sampling methods can be used to try to minimise the number of model runs required to give an accurate distribution for the output quantity. The best-known example of a sampling method is probably Monte Carlo simulation. The methods are particularly suitable for continuous models (and particularly finite element models) because they do not require the statistical distributions to be propagated through the equations explicitly, and the equations for a continuous model are generally too complex for explicit propagation to be possible. The main drawback with the methods is that the repeated runs of the model can be time-consuming.

Stochastic techniques treat the uncertain input quantities as distributed random variables and develop stochastic partial differential equations to describe the distribution of the results. These methods are complicated even for simple one-dimensional cases, and can become even more so as more parameters are regarded as uncertain and more dimensions are included. However, they are still useful, particularly for cases where a proprietary package is not being used and so derivative information may be available to the modeller.

In addition to these methods, there are techniques which identify problematic parameter values. It is possible that some parameter values can lead to unstable or non-unique solutions of a problem. These parameter values (sometimes known as bifurcation points) can be extremely troublesome when calculating uncertainties. Solutions may be extremely sensitive to choices of parameter value around a bifurcation point, and this sensitivity can lead to a wide spread in the values of a result of interest, thus giving a very large uncertainty. For more detail, see section 4.2 on spurious solutions.

It is possible that different numerical methods will be affected to different degrees by parameter uncertainties, and clearly if it is possible to choose a method that is not sensitive to parameter error then this should be done unless the modeller has reason to believe that the sensitivity is a feature of the physical system too.

3.5 Linear algebra error

Linear algebra errors are considered as a software error rather than a modelling one for several reasons. One reason is that such errors are generally beyond the control of users of proprietary software, although some packages do allow the user to choose the solution method to help reduce the errors, and careful choice of a system of units for a model can make errors less likely. For instance, a system of units that avoids calculation of very large stresses by giving material properties in terms of megapascals instead of pascals may be less susceptible to numerical errors. Another reason is that there is a large number of library software packages offering best-practice algorithms for solutions of large systems of linear and non-linear equations, and many of the packages provide error estimates as part of their calculations, so such information is already available to users writing their own software. A methodology for testing continuous modelling software to identify linear algebra errors is described in the report "Testing Continuous Modelling Software" [25].

Iterative methods are particularly common in the solution of linear algebra problems generated by continuous models. The methods are used is because the solution procedures often involve very large sparse matrices. The size is due to the level of detail required for accuracy, and the sparseness is because in general the equations relate small groups of nodes of the mesh that are close together, which means that each row of the matrix only has a small number of entries. In general, iterative techniques are not terminated until they reach a solution which gives residues smaller than some user-selected tolerance when substituted back into the problem. This termination procedure means that often the user has a good degree of control over the linear algebra errors, although the methods may not terminate in a satisfactory manner if the matrix is ill-conditioned.

This report will not go into depth about the causes of linear algebra error, and the interested reader is referred to the deliverables of the project mentioned above for ways of testing for the existence and magnitude of such errors.

4 Overview of validation

This section aims to give a general overview of points to take into consideration when validating a model. Many of the points are standard modelling practice such as the importance of validation during the development stages, but it is necessary that they are understood. This section also includes information on multiple and spurious solutions, which can be a cause of incorrect validation results.

4.1 Validation strategy

Mathematical models are often used to investigate general trends qualitatively to save on experimental costs, and so detailed examination of numerical results may not be required. The objectives of the modelling activity will strongly affect the choice of validation methods and the manner in which they are applied. The modeller will have considered the purpose of the model during development, and will have identified whether the final aim is to gain a qualitative understanding of some phenomenon of interest, to produce the simplest possible model that incorporates the major features of the process of interest, or to produce a comprehensive descriptive model of the system. These possibilities can require an increasing amount of validation, but it may be worth validating to a degree higher than necessary to ensure that erroneous conclusions are not reached about the reasons for the obtained results.

Degree of validation

Fowkes and Mahony [5] consider some examples where the degree of validation has been chosen to match the purpose of the model. In one such example, Volterra introduced models to explain the changes in the shark population in the Mediterranean qualitatively in circumstances in which it would have been impossible to obtain reliable quantitative data. His work led to useful insights into population growth. Another useful example [5] shows the importance of re-validating models when their complexity increases: early attempts to understand weather patterns, by including compressibility effects in meteorological models, led to predictions that weather fronts would move at the speed of sound, but omitting such effects produced better predictions of weather front behaviour.

By their nature, continuous models produce results over the whole of the computational domain. Often most of the results in most of the domain are not of interest, since the physical process that has been modelled may only produce results in one region. Similarly, many models produce predictions of quantities that are not of direct relevance to the experiment. For example, a finite element model of an adhesive joint in tension will generate stress and displacement results within the adhesive and the adherends, but it could be that the only experimental measurements made are of the force required to deform the joint, which means that most of the model results cannot be verified by the experimental data. This partial validation means that care must be taken to consider which parts of the model have been validated when looking at the results. In the case above, it may not be reasonable to draw conclusions about the stresses within the joint as no experimental stress data exists.

Some problems may have non-unique solutions and it is difficult to be certain that the model provides the solution corresponding to the physical situation being modelled. This concept is further discussed in section 4.2. Often the physical problem has a unique solution, but the mathematical formulation of the problem may have multiple solutions, particularly if simplifying assumptions have been made. The likelihood of hysteresis and multiple solutions should be considered during validation, particularly if an otherwise well-behaved model produces unexpected results for some parameter values but not others. This is discussed further in section 4.2.

Validation during model development

Validation should be a part of the model development process as well as a final test of the model's suitability for use. As each new level of complexity is added to a model, the addition should be checked for internal consistency, and where possible external consistency as well. External consistency checks of the whole model should probably be saved until it is almost completed, but if alternative data is available to check a sub-component of the model this should be used.

For example, suppose a material model has been developed and implemented in a finite element software to describe plastic deformation. Eventually the model will be used to describe deformation of objects that are complicated shapes, but before the model is used on large simulations requiring many elements, it is tested on a series of simple geometric elements and load-cases to ensure that it behaves correctly under all loading conditions and when applied to every different element type with which it is likely to be used. The model results are compared to experiments that produce simple stress states in the material. This process is effectively an external consistency check of the material model. The overall validation process for the model is shown in figure 2. This validation strategy of first solving simple problems and then building up to more complex ones is similar in approach to a methodology developed for testing continuous modelling software [25].

Detailed examination of the results of a validation can often give useful information as to how the model can be improved or corrected, particularly when the results include a distribution of error estimates over space or time. Large errors in one region of a model can indicate that further mesh refinement is needed there, or they can indicate that the definition of a boundary condition needs improvement. Similarly, errors that grow over time can indicate that the method of solution does not have ideal stability characteristics for the problem of interest.

Model validation and error estimation can be built into the implementation of the model itself. Many modern finite element packages offer adaptive meshing and adaptive time stepping, where an error estimate is generated based on the calculated solution, and the mesh density or time step size is altered to keep the size of the error within acceptable bounds. Examples of some of the techniques that are used to generate the error estimates are described in section 7.2 on a posteriori methods.

Wherever possible, the data used to generate a model's input parameters and the data used to validate the model should be different. If the input parameters were calculated from the validation data, the validation shows that the chosen model describes the validation data and that the parameters have been calculated correctly. This aspect is described further in section 5.3 on comparison with experimental data.

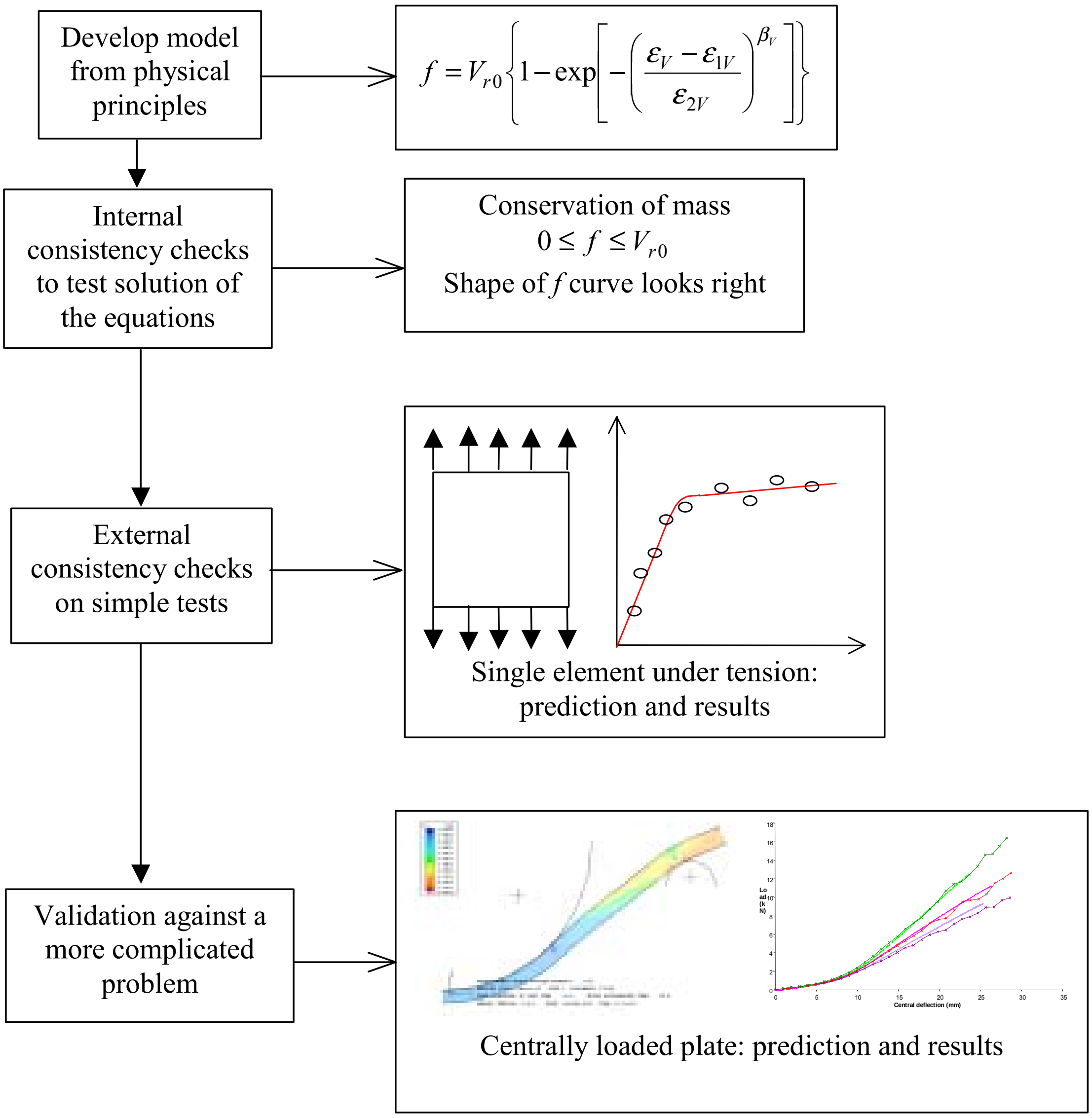
Develop model from physical principles
$f = V_{r0}\left\{1-\exp\left[-\left(\frac{\varepsilon_V-\varepsilon_{1V}}{\varepsilon_{2V}}\right)^{\beta_V}\right]\right\}$
Internal consistency checks to test solution of the equations
Conservation of mass
$0 \le f \le V_{r0}$
Shape of f curve looks right
External consistency checks on simple tests
Single element under tension: prediction and results
Validation against a more complicated problem
Central deflection (mm)
0
5
10
15
20
25
30
35
0
2
4
6
8
10
12
14
16
18
Centrally loaded plate: prediction and results

Figure 2: Flowchart showing a typical development and validation process. The boxes on the left are the typical steps in the process, and those on the right show how they are implemented for the material model mentioned above. Here, *f* is the dependent variable, ε_V is an independent variable, and the other terms are material parameters.

Some methods may need to be re-applied every time a model is used with new parameters. For instance, altering the material properties in a model by a large amount is likely to require re-checking of the mesh convergence and time step stability, but it may not need re-checking against a known solution if the calculation method has already been validated that way. In the most severe cases, the entire model may need re-designing. For example, the physics governing behaviour from quantum level to bulk continuum properties requires a range of different mathematical models to describe it, so radical changes in length scale will invalidate a model and necessitate a completely new model.

Wherever possible, records should be kept of validation, in a similar way to testing records generated during software development. The information recorded should include the method used, input data, expected result, and obtained result. This information will make it easier to identify possible error sources in any future model development, and will make it clearer which aspects of the model have been tested. As the use of computer models in metrology increases, records like this will become useful for ensuring reliability and traceability of results.

4.2 Multiple and spurious solutions

Multiplicity of possible solutions occurs in many physical situations, and is commonly seen as a hysteresis phenomenon. Hysteresis is a general term for any system where its response to a load depends upon the past history of the system as well as its current state. It occurs in cyclic loading in which a component is loaded and unloaded repeatedly. Common examples include unloading and reloading of a material that has passed its yield strength, and magnetization of a ferromagnetic material by a time-varying field. A typical response curve with hysteresis is shown in figure 3. Some values of the load correspond to multiple values of the response.

Generally, experimental repeatability indicates that the physical system corresponds to a locally stable solution but there may be more than one locally stable solution. For example, the motion of a vibrating object could be dominated by any of its fundamental modes and the solution mode realised by the system in a given situation may depend sensitively on the experimental factors.

As well as genuine multiple solutions, spurious multiple solutions can be caused by some numerical methods. The existence of these multiple solutions can lead to problems with validation, particularly when comparing with other methods or with experimental data. If the model has provided one of several solutions, but the experiment or alternative method has calculated a different one, a correct model could easily be rejected.

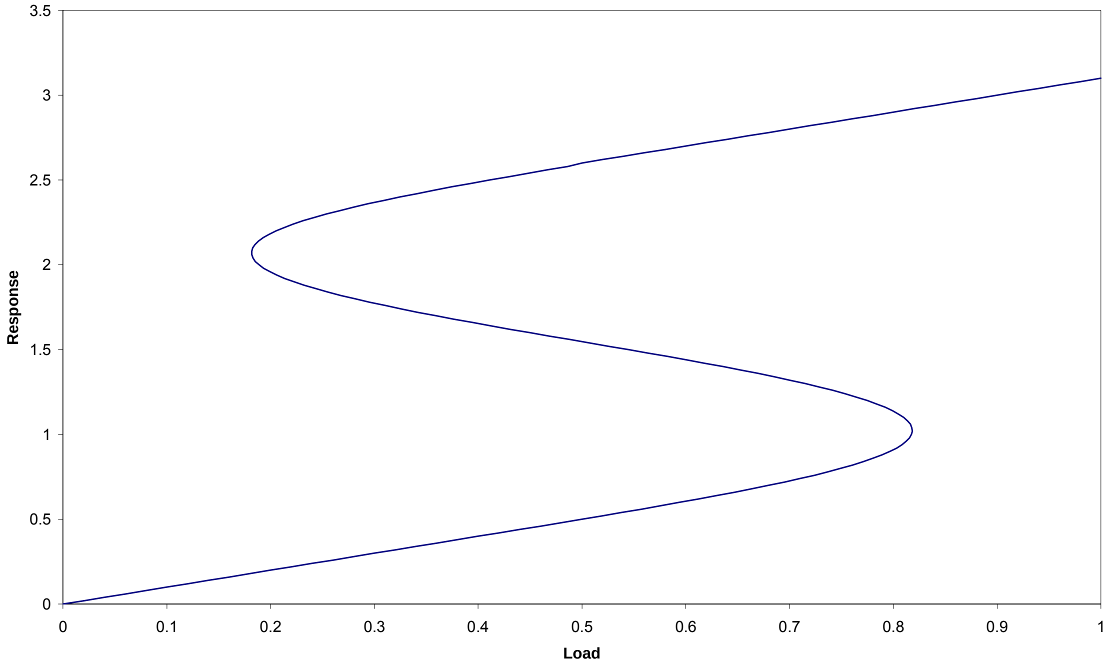

Figure 3: Example of a response curve with hysteresis.

One of the simplest examples to consider is the long-time solutions of a system of differential equations. Consider a set of differential equations

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x};\boldsymbol{\lambda}), \qquad (6)$$

where **x** are variables and **λ** are parameters. The system has a fixed point if there is some set of variable values $\mathbf{x}_f$ such that $\mathbf{F}(\mathbf{x}_f\ ;\ \boldsymbol{\lambda}) = \mathbf{0}$, so that once the solution **x**(t) reaches $\mathbf{x}_f$ it remains there. The system has a periodic orbit if there is a solution and an associated time *T* such that $\mathbf{x}_p(t + T) = \mathbf{x}_p(t)$. Again, the existence of a periodic orbit means that once the solution reaches this periodic orbit it will stay within that path. Both of these behaviours are likely long-time solutions for physical systems. For example, most thermal systems tend towards a fixed point of uniform temperature, and periodic solutions occur in many dynamical systems. All of the following examples are based on systems of the form (6), but the key points illustrated apply to partial differential systems as well.

Bifurcation points

As the equation (6) states, these solutions will depend on the values of the parameters **λ**, and in many cases it is possible for more than one such solution to exist simultaneously for some values of the **λ**. The existence of these solutions is a property of the mathematical description and possibly the physical system, rather than the method used to solve the problem. As the **λ** vary, the stability of the solutions will vary as well. Unstable solutions are such that if the initial conditions are perturbed slightly, the solution will move away from the fixed point, and stable ones are such that the solution will still reach the fixed point if the perturbation is small. The values of **λ** at which the stability of solutions changes, and at which new fixed points or periodic solutions are created, are called **bifurcation points**. There are various different types of bifurcations depending on what changes in stability occur.

As a simple illustration of the coexistence of multiple solutions consider a system that is a model of a chemical reaction:

$$\begin{aligned} u_1' &= -u_1 + \lambda(1-u_1)\,e^{u_2}, \\ u_2' &= -3u_2 + 14\lambda(1-u_1)\,e^{u_2}. \end{aligned} \qquad (7)$$

When the parameter $\lambda = 0$, this system has a fixed point $u_1 = u_2 = 0$. Figure 4 shows what happens to the fixed point as λ is varied, plotting the l_2 norm of **u** against λ. This system also has periodic solutions that arise through a Hopf bifurcation point at $\lambda \approx 0.131$. A bifurcation that produces a periodic solution from a fixed point is called a Hopf bifurcation. The path of the periodic solutions is indicated on the figure by the circles.

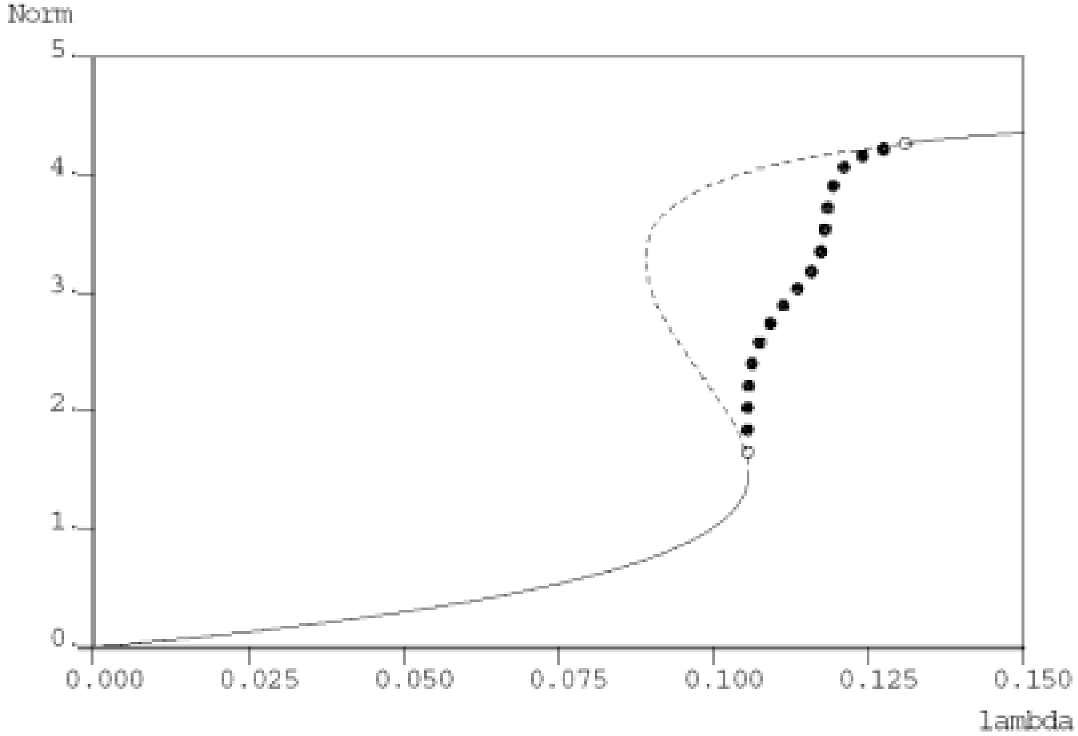

Figure 4: Bifurcation diagram for the system (7) as λ varies. The continuous line is a stable fixed point, and the broken line shows the unstable fixed point. The path of the periodic solutions from the bifurcation are plotted as filled circles and the bifurcation points are shown as open circles.

So for example at $\lambda = 0.100$ there are three possible fixed point solutions, one stable and two unstable. At $\lambda = 0.110$ there are two solutions, one an unstable fixed point and the other a stable periodic solution.

Viewing the system as a dynamical system allows us to understand the structure of the solutions as the system parameters vary. The existence of bifurcations may be important when there is large uncertainty in the parameters.

For the coexistence of stable stationary solutions, consider the following example from enzyme modelling:

$$\begin{aligned} u_1' &= (\lambda - u_1) + (u_2 - u_1) - 100R(u_1), \\ u_2' &= (\lambda - u_2) + (u_1 - u_2) - 100R(u_2), \\ R(u) &= \frac{u}{1 + u + u^2}. \end{aligned} \tag{8}$$

The solution of this system gives a classic hysteresis curve with a secondary bifurcation as shown in figure 5.

For example, for a fixed value of $\lambda = 26$, starting from different initial data we can find any of the stable fixed points in the system. Which solution is correct corresponds to the particular physical situation, and the basin of attractions for different solutions can be very complicated (even fractal: further discussion of this problem is available [13]). To complicate matters further, there may be phenomena present in the physical situation, but absent from the model, such that a physically stable solution corresponds to an unstable solution of the model. Thus, instability is not necessarily a reason for rejecting a computed solution.

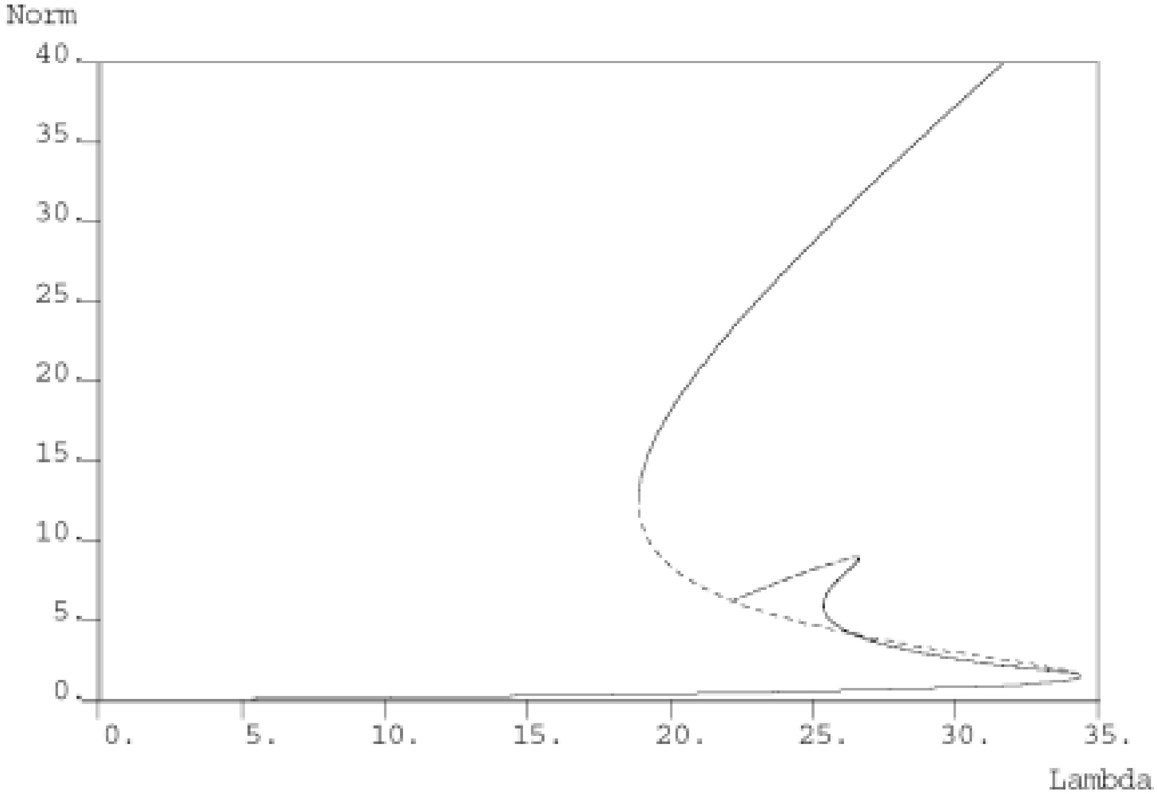

Figure 5: Bifurcation diagram for the system (8) as λ varies. The continuous line denotes a stable fixed point and the broken line an unstable fixed point. Note the coexistence of multiple (locally) stable fixed points.

In general, because in simple terms stable solutions attract and unstable ones repel, it is very difficult to find the unstable solution using normal numerical integration methods for systems of ordinary differential equations such as the forward Euler method. The two examples above were computed using a numerical continuation method [15].

Spurious solutions

In examples (7) and (8), multiple solutions exist because they are a feature of the differential equations of the model. However, it is possible for numerical schemes to introduce spurious solutions during the discretisation process. As was stated in section 2.3, the usual proofs of convergence only apply to discretised models over a finite time period, and so the standard methods will not give a reliable error estimate for long-time solutions. This issue has been addressed and the long-time stability and convergence properties of various numerical methods investigated [14], starting from a dynamical systems viewpoint.

As an example of spurious multiple solutions, consider the ordinary differential equation (ODE)

$$u' = \frac{-\lambda u}{1+u^2} \tag{9}$$

and apply the explicit two-stage Runge-Kutta method for $u' = f(u)$: $U^{n+1} = U^n + hf\left(U^n + hfU^n\right)$.

The differential equation has a single fixed point solution (u = 0). However, if $h > 1/\lambda$ the Runge-Kutta method has three fixed points:

$$U = 0, \quad U = \pm\sqrt{\lambda h - 1}. \qquad (10)$$

U = 0 is also a fixed point of the ODE and so should be produced by the method, whereas the other two fixed points are spurious fixed points and are not a feature of the continuous equations. It should be noted that solutions that converge to spurious fixed points are often smooth and so it is not always evident that the fixed point is spurious. Spurious periodic solutions are also possible. Figure 6 shows the status of the positive spurious steady solutions for λ = 500 for a variety of step sizes h. Initially there is no spurious solution until h = 0.002, at which point the spurious solution exists and is stable, and the genuine steady-state solution becomes unstable. Another bifurcation exists at h = 0.004, where the spurious solution in (10) becomes unstable and spurious solutions of period 2 come into existence. A solution of period 2 is a solution that oscillates between two fixed values, the values in this example being

$$\sqrt{h\lambda/2} + \sqrt{h\lambda/2 - 1} \quad \text{and} \quad \sqrt{h\lambda/2} - \sqrt{h\lambda/2 - 1}$$

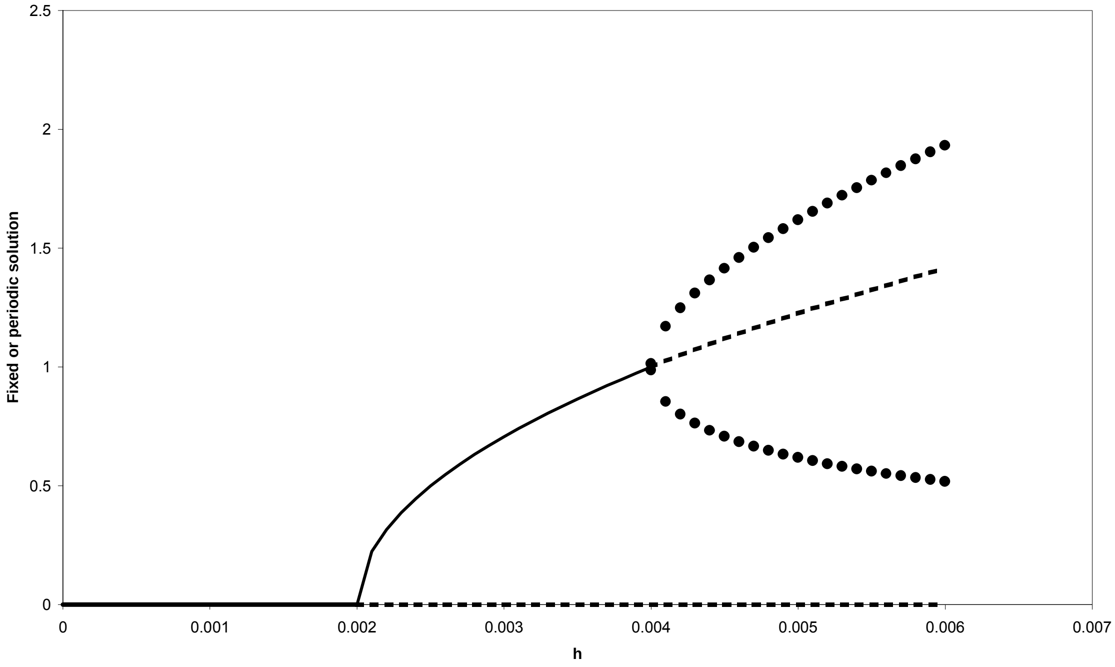

Figure 6: Variation of steady state solutions to problem (9), with step size h for λ = 500. A continuous line denotes a stable solution, filled circles denote a solution which of period 2, and the broken line is an unstable solution.

In general, spurious fixed points can be detected by reducing the step size. Doing so usually causes the spurious fixed point either to converge to a true fixed point of the system or to become unbounded and grow. An example of this effect is the application of the perturbed Euler method,

$$U_{n+1} = U_n + hf(U_n) + h^2(1 + U_n^2)$$

to $u' = f(u) = u$. This system has fixed points at

$U_n^{\pm} = \left(-1 \pm \sqrt{1 - 4h^2}\right)/2h$, so as $h \to 0$, $U_n^{+} \approx -h \to 0$ and $U_n^{-} \approx -1/h \to -\infty$.

Thus one spurious solution tends towards the correct fixed point and the other grows without bound. Using a variable time stepping algorithm will often avoid spurious fixed points and periodic solutions altogether. The spurious fixed points that are features of the numerical scheme generally have their values determined by the time step (as shown in the examples above), whereas a real fixed point of the differential equations has the same value for all step sizes. Hence if the time step changes, a spurious fixed point will no longer be a fixed point and the solution will progress correctly.

4.3 Validation methods

The next three sections provide practical guidance on validating continuous models. Methods in the same section can be considered to be the same type of consistency check. However, these definitions are not clear-cut: for instance, validation using conservation laws could be considered to be an inter-model consistency check if it is known that the differential equation has a conservation property and the test is whether or not the discretised version preserves this property. This partitioning is also useful on the grounds of ease of application and familiarity. In general, the external consistency checks are familiar to most metrologists and are easiest to apply, the internal consistency checks are familiar to experienced modellers and are reasonably straightforward provided the necessary alternative solutions exist, and the inter-model consistency checks are not commonly used directly and require careful application.

The methods described in this section are not mutually exclusive. Ideally, validation should include as many of them as possible, since often one will check for a problem that another may miss. Several could be regarded as testing methods rather than error estimations as they give grounds for rejecting a model and indicate fatal errors rather than estimating the size of an error that can be reduced but not eliminated altogether. Generally, the former type are qualitative criteria and the latter are quantitative, and thus can be used for ranking models if more than one formulation is available.

Each of the sections describing a method provides

- a description of the method
- a statement of the aspect of the model that is being validated
- an example of a successful application of the method
- warnings about the drawbacks of the method
- an example of the insufficiency method.

Examples are drawn from several different areas of physics.

5 Internal consistency checks

Internal consistency checks test the consistency of the particular implementation of the model. They usually involve checking against reference solutions, either from analytic solutions or from reference software, which makes them similar to some software testing methods [26].

5.1 Comparison with analytic solution

Comparison with an analytic solution to a problem is an ideal internal consistency check. It checks that the software solves the chosen equations, it gives an indication of the computational errors being generated during calculation, and it enables comparison of alternative calculation methods for the same problem.

For example, consider a coated substrate as shown in figure 7. The coating has a crack through its thickness and is partially debonded from the substrate. An analytic model has been developed that predicts the stress in Region I of figure 7, and the rate of change of the potential energy of the system during a finite increment of debond growth, under tensile and bending boundary conditions. These quantities can also be calculated using finite element analysis, although care must be taken to use a fine mesh around the crack tip, and the energy release rate must be calculated using a surface integral, which can be problematic. Figure 8 shows the mesh that was used, with elements used for the integral calculations shaded heavily.

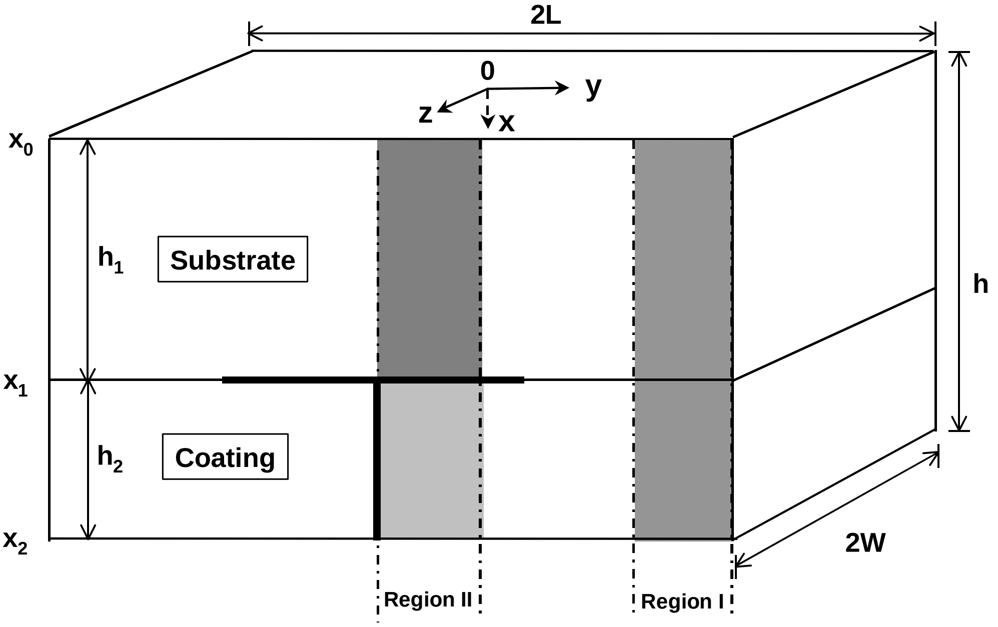

Figure 7: Diagram showing substrate and coating with through-thickness crack and partial debond.

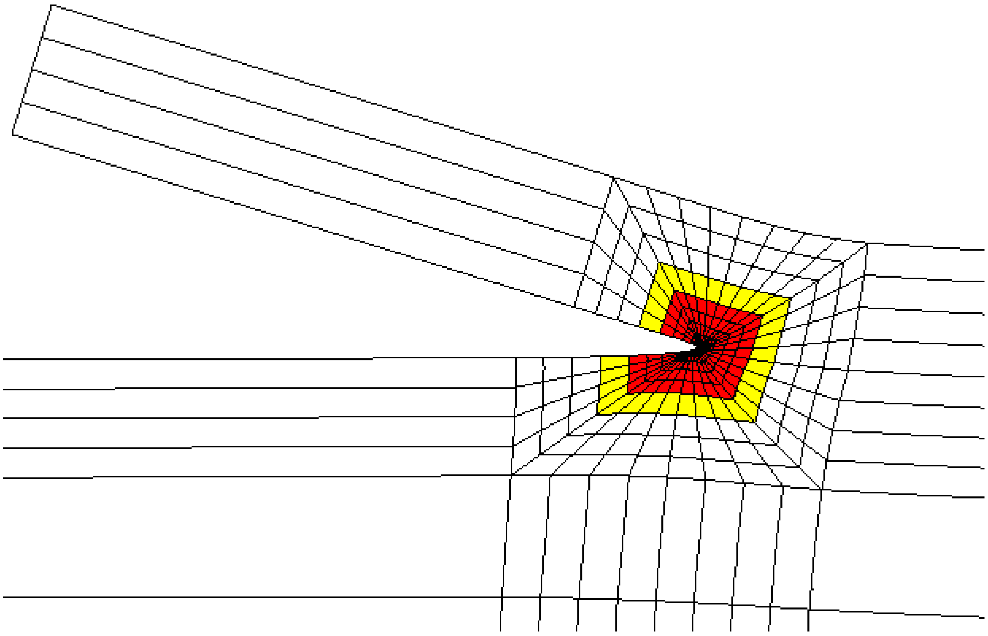

Figure 8: Diagram showing the mesh used around the debond tip. The shaded elements were used to calculate the energy release rate.

Both models were solved with "tension" and "tension with bending" boundary conditions to generate comparable data. Two sets of results were compared in the validation. The axial stresses in the coating and substrate in Region I are shown in table 1 for the two sets of boundary conditions. The rate of change of potential energy was calculated for a number of different debond lengths by the FE model. The analytic model assumes that there is no interaction between regions I and II, so that the rate of change of potential energy is independent of the debond length and only one value was calculated for each set of boundary conditions. Ideally the FE results would converge to the analytic model results as the debond length increases. The results of these calculations are shown in figures 9 and 10.

	Boundary conditions	**FE Model**	**Analytic model**
Axial stress in coating (GPa)	Tension	43.795	44.271
Axial stress in substrate (GPa)	Tension	22.955	22.928
Axial stress in coating at free surface (GPa)	Bending & tension	83.115	84.115
Axial stress in coating at interface (GPa)	Bending & tension	83.380	84.380
Axial stress in substrate at free surface (GPa)	Bending & tension	2.344	2.293
Axial stress in substrate at interface (GPa)	Bending & tension	43.029	43.564

Table 1: Stress results from the finite element and analytic models.

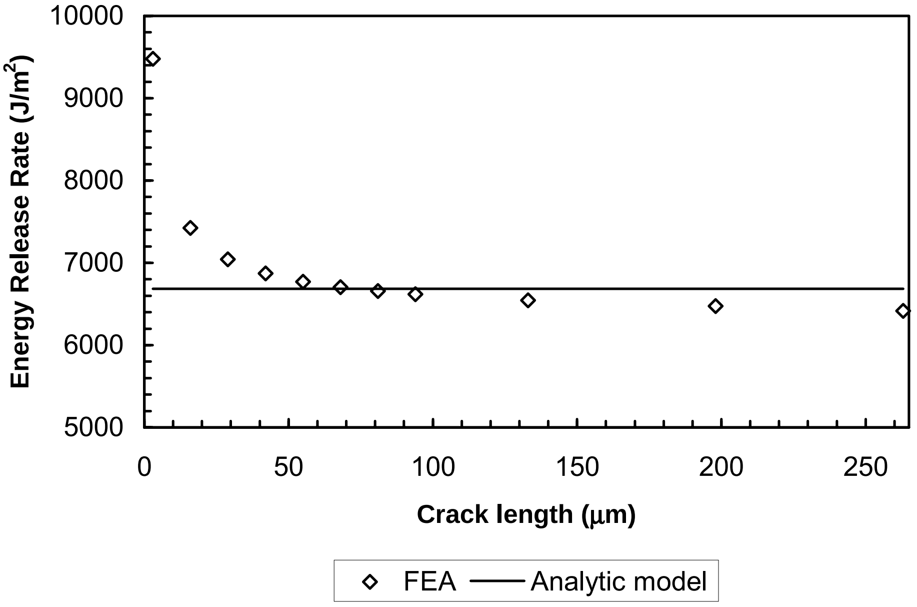

Figure 9: Comparison of finite element and analytic model results for energy release rates under tension conditions.

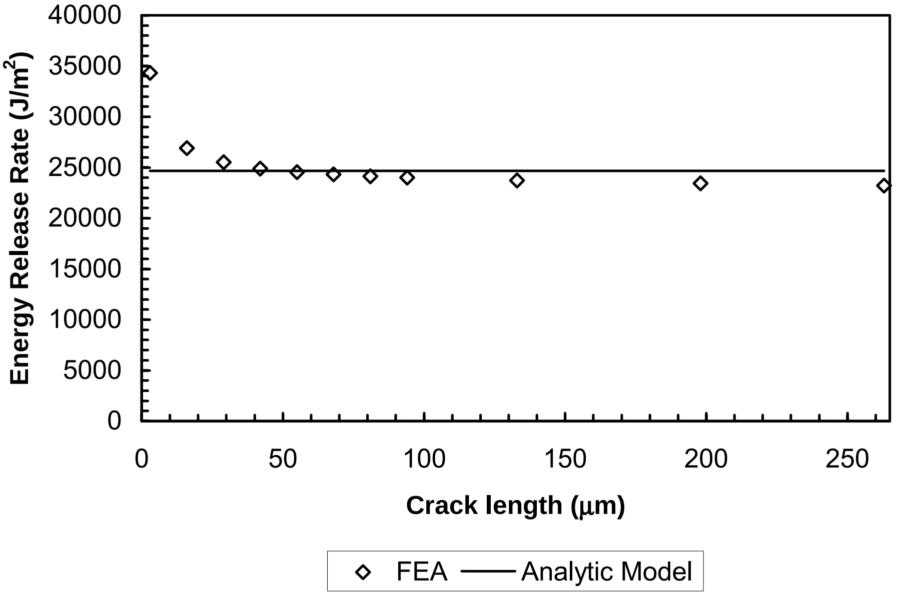

Figure 10: Comparison of finite element and analytic model results for energy release rates under bending and tension conditions.

The agreement between the results of the FE model and the analytic model is generally quite good, but not perfect. The reasons for the discrepancies are discussed in detail later in this section, but they are likely to be due to the different assumptions made when generating the two models.

In some cases an analytic solution will exist for a particular region of a model, for instance, a symmetry line or along $x = 0$, and a combination of comparison with this solution and checking that the results in the rest of the solution domain “look right” (see section 6.1) can be useful.

The main drawback to the method is that very few analytic solutions exist for anything but the simplest of problems. If most problems had analytic solutions then discrete approximation methods for partial differential equations (PDEs) would not be so important. However, a scarcity of analytic solutions does not mean that none exist, and those that do can be used to check the software, since if the software cannot solve problems with analytic solutions correctly it is unlikely to be able to solve other problems.

Another problem is that it can be difficult to reproduce the exact assumptions that go into an analytic solution, particularly in a finite element or finite difference model. Often assumptions are made about quantities over which the modeller has little control.

For example, consider a simple beam as shown in figure 11. An analytic solution can be derived for σ_{xx}, σ_{zz}, τ_{xy}, u, v (as defined in section 2.1), provided that the assumptions $\sigma_{yy} = \tau_{xz} = \tau_{yz} = w = 0$ are made. The boundary conditions for the analytic solution are that $u(0, y, z) = 0$ and $v(0, 0, z) = 0$. If this situation is modelled using finite elements, it is simple to ensure that the conditions on the displacement are met by using a plane strain model and fixing the appropriate nodes, but it is not possible to ensure that $\sigma_{yy} = \tau_{xz} = \tau_{yz} = 0$ throughout the beam.

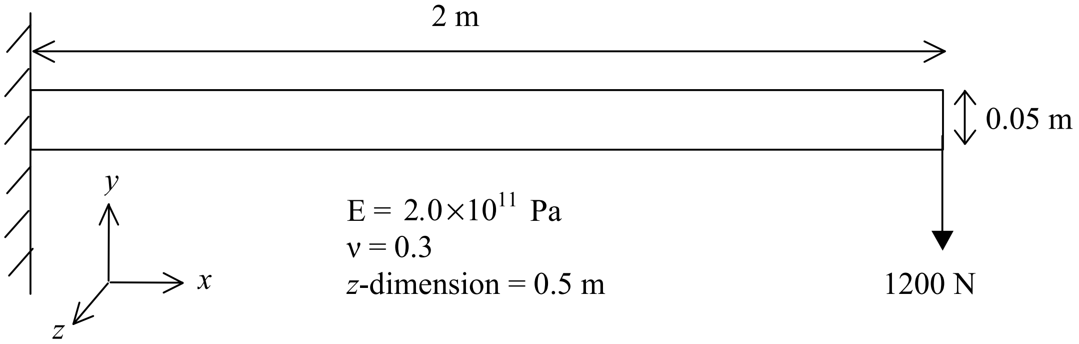

Figure 11: Beam, fixed at the left-hand end, with the y-dimension much smaller than the x and z dimensions, so that a plane strain analysis can be used.

Figure 12 shows the calculated results for σ_{yy} along the lower surface of the beam. Whilst the values are smaller than those calculated for σ_{xx} and σ_{zz} (maximal values 11.4 MPa and 3.4 MPa respectively), they are comparable with the values for τ_{xy} at the ends of the beam. This means that care must be taken when validating the model against the analytic solution, since the two are not solving the same problem.

Similarly, in the example given above, assumptions were made in the analytic model that it is impossible to reproduce in the finite element model (such as stresses being zero on free surfaces and there being no interaction between regions I and II). It is likely that these differences contribute to the discrepancy in the results. This means that care must be taken to consider all possible differences, to eliminate as many as possible, and to consider how any that cannot be eliminated may have affected the results.

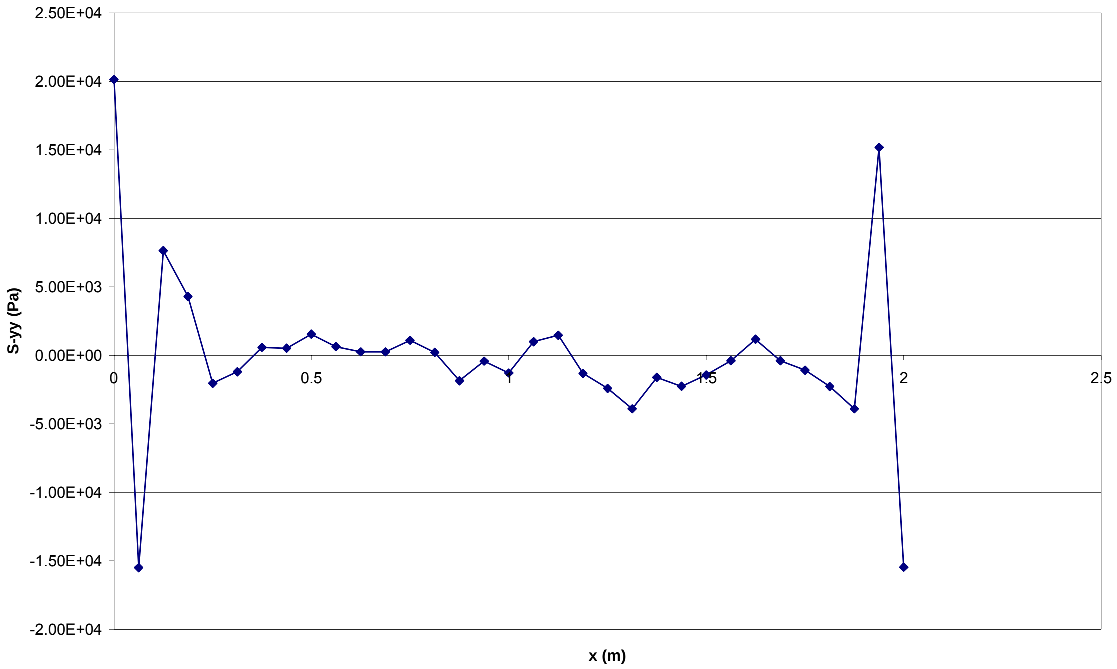

Figure 12: Calculated σ_{yy} results for a finite element model of the beam. The analytic solution assumes that $\sigma_{yy} = 0$ everywhere.

5.2 Comparison with equivalent models

There are two types of equivalent problems that may be useful for validation purposes. The first type is problems that are simplifications of the model being validated, and the second type is problems that are mathematically equivalent to the model in question.

As with other checks involving comparisons, one of the main drawbacks with this method is the requirement that the equivalent model needs to have been independently validated for any comparison with it to be meaningful (since otherwise they could both be making the same mistake).

5.2.1 Simplified models

Comparison with a simplified model is similar in outlook to the examination of extreme conditions suggested in section 6.1 of this report, as it takes a case of the model that is easier to understand and uses it to check results.

Simplified versions of a model are often developed before a more complex version so that some experience of the problem and solution technique can be gained before the more difficult task is started. Common simplifications include:

- assumption of reflectional symmetry, rotational symmetry, or axisymmetry to reduce problem size,
- assumption of linearity of material properties or boundary conditions to reduce complexity of the solution method,
- assumption of constant properties or conditions to reduce a dynamic problem to a static one.

More details on these simplifications are given in the "Guide to the use of finite element and finite difference software" [25]. In many cases, by careful specification of the model inputs the more complex model can be used to generate results that should be identical to those from the simplified model. So for example, if all conditions and properties are specified to be fully symmetric, a three-dimensional model should produce the same results as a model that solves a two-dimensional axisymmetric problem. Similarly, the solution to a dynamic heat-flow problem at large time values may converge to that of an equivalent static problem.

For example, consider two common models for plasticity in solids. The Gurson model [18] is for solids that can include voids, with yield function

$$\frac{\sigma_e^2}{\sigma_M^2} + 2f \cosh\left(\frac{\sigma_K}{2\sigma_M}\right) + f^2 - 1 = 0, \tag{11}$$

where σ_M is the yield stress of the material with no voids, $\sigma_K = \sigma_{11} + \sigma_{22} + \sigma_{33}$ is the pressure stress, $\sigma_e = \sqrt{(S_{ij}\, S_{ij} / 2)}$ where $S_{ij} = \sigma_{ij} - \delta_{ij}\, \sigma_K / 3$ is related to the equivalent deviatoric stress, and f is the volume fraction of voids in the material. This model will also include an equation defining void nucleation and evolution. The von Mises model, probably the most common plasticity model, has yield function

$$\frac{\sigma_e^2}{\sigma_M^2} - 1 = 0\,. \tag{12}$$

It is clear that a model using the Gurson model and ensuring that void nucleation never occurs so that f is always zero should give the same results as the von Mises model. To test this property, the two models were run under uniaxial tension conditions. The results for axial stress versus axial strain are shown in figure 13.

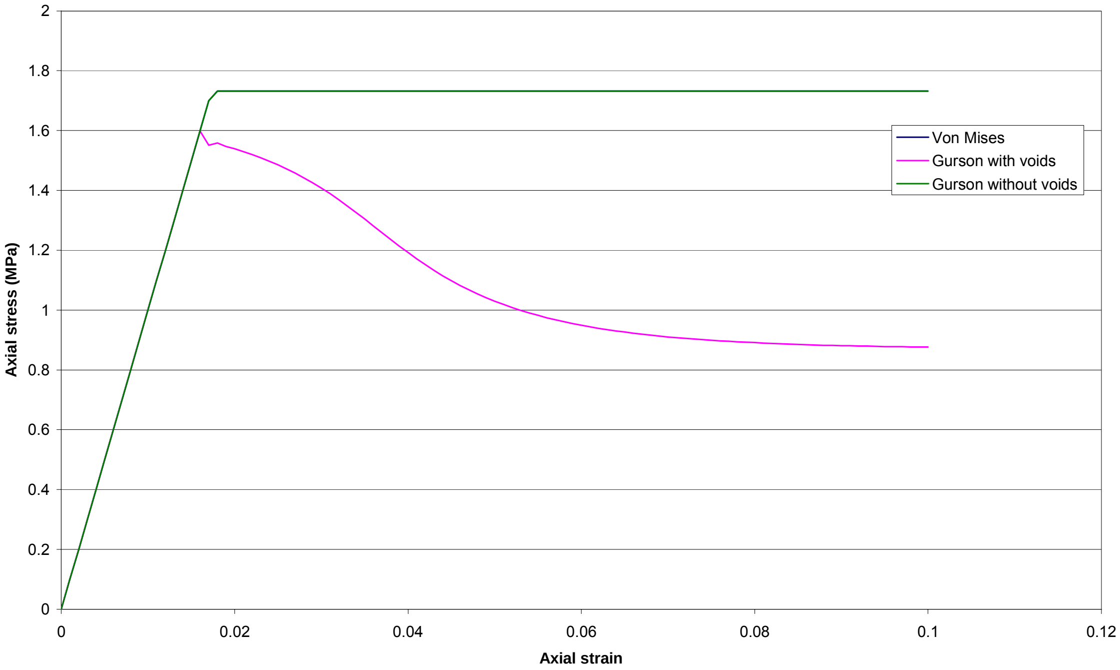

Figure 13: Plot of axial strain versus axial stress for the von Mises model and the Gurson model with and without voids. The results of the Von Mises model and the Gurson model with no voids are identical.

The results of the von Mises model and the Gurson model without cavities are identical: calculation of the differences over time gave a zero result throughout, despite the use of a Newton-Raphson solver to find the solution to (11). This is partly due to the simplicity of the geometry and loading conditions: a more complicated model could result in larger rounding errors and accumulation of error over time.

The main drawback of comparison with a simplified model is that it does not validate the part of the model that has been removed in the simplification. For instance, in the example above there is no guarantee that the terms in (11) involving f are calculated correctly.

5.2.2 Mathematically equivalent models

There are several examples of different physical systems described by mathematically equivalent models. Under many conditions, the transfer of heat and matter can both be described by the convection-diffusion equation

$$\nabla.(a\nabla u - \mathbf{b}u) + cu + d = e\frac{\partial u}{\partial t},$$

where u is the variable of interest, and a, $\mathbf{b}$, c, d, and e can all be functions of position or time. The electric potential of electromagnetic propagating waves in the absence of charge obey the Helmholtz equation

$$\nabla^2\varphi + k^2\varphi = 0,$$

as does the time-independent velocity potential of periodic acoustic waves. Many mechanical systems of masses, springs and dampers are described by the same ordinary differential equations as electrical circuits containing inductors, capacitors and resistors (see below for an example).

Similarly, many problems can be solved for a single value of a non-dimensional parameter, which could represent a wide variety of physical situations. For example, the equations for steady viscous fluid flow can be made non-dimensional:

$$\nabla.\mathbf{u} = 0,$$

$$\mathbf{u}.\nabla\mathbf{u} = \nabla p + \frac{1}{Re}\nabla^2\mathbf{u},$$

where Re is the Reynolds number, **u** the fluid velocity, and p the pressure. $Re = \eta/L\rho U$, where L is a typical length, U a typical velocity, ρ the fluid density, and η the fluid dynamic viscosity. This consideration of the equations in a non-dimensional form means that a slow flow in a thin fluid may have the same solution as a faster flow in a thicker one. Similarly, in acoustics, problems with the same value of the dimensionless parameter ka can be compared.

This mathematical equivalence means that solutions to the two problems can be compared for validation purposes, provided that the correct non-dimensionalisation of all results has taken place to ensure that the comparison is valid.

As an example, consider the charge in an electrical circuit containing a charged capacitor C, a resistor R and an inductor L in series (figure 14a). Assume the initial charge in the capacitor at time 0 is Q_0, the initial current in the circuit is zero, and consider the situation at time t. Kirchhoff's second law states that the total voltage across a set of passive components is always equal and opposite to the source voltage. Therefore, the sum of the voltage differences across all the circuit elements (including the source) is always zero. In this case there is no source voltage, so if the voltages are denoted as V,

$$V_C + V_R + V_L = 0. \tag{13}$$

Expressions for the individual potential differences across the components in terms of the current I through them and their properties are

$$V_C = \frac{1}{C}\int I dt, \qquad V_R = IR, \qquad V_L = L\frac{dI}{dt}. \tag{14}$$

Hence, writing current as the rate of change of charge, so that $I = dQ/dt$, and substituting (14) into (13),

$$L\frac{d^2Q}{dt^2} + R\frac{dQ}{dt} + \frac{Q}{C} = 0,$$
$$Q(0) = Q_0, \quad \left.\frac{dQ}{dt}\right|_{t=0} = 0.$$

Now, to put the equations into non-dimensional form, write

$Q = Q_0\hat{Q}$ and $t = \hat{t}\sqrt{LC}$,

as these choices are consistent with the units of the various quantities. Rearranging gives

$$\begin{aligned} &\frac{d^2\hat{Q}}{d\hat{t}^2} + \frac{R\sqrt{C}}{\sqrt{L}}\frac{d\hat{Q}}{d\hat{t}} + \hat{Q} = 0, \\ &\hat{Q}(0) = 1, \quad \left.\frac{d\hat{Q}}{d\hat{t}}\right|_{\hat{t}=0} = 0. \end{aligned} \tag{15}$$

Now consider a pendulum of mass m on a massless string of length l undergoing small oscillations in a viscous medium. Assume that the pendulum is initially at rest and released from an angle θ_0. The forces acting on the pendulum are due to viscous drag (taken to be proportional to the velocity of the pendulum, with constant of proportionality D) and gravity (figure 14b). Resolving these forces in the direction of motion of the pendulum and applying Newton's second law gives

$$\begin{aligned} &ml\frac{d^2\theta}{dt^2} + Dl\frac{d\theta}{dt} + mg\sin\theta = 0, \\ &\theta(0) = \theta_0, \quad \left.\frac{d\theta}{dt}\right|_{t=0} = 0. \end{aligned} \tag{16}$$

If θ_0 is sufficiently small and D is positive, the oscillations will remain small due to the viscous damping, so that $\sin\theta \approx \theta$. Rewriting the equations in a non-dimensional form by writing

$\theta = \hat{\theta}\theta_0$ and $t = \hat{t}\sqrt{l/g}$,

and inserting these expressions into the linearised form of (16) gives

$$\begin{aligned}&\frac{d^2\hat{\theta}}{d\hat{t}^2}+\frac{D\sqrt{l}}{m\sqrt{g}}\frac{d\hat{\theta}}{d\hat{t}}+\hat{\theta}=0,\\&\hat{\theta}(0)=1,\quad \left.\frac{d\hat{\theta}}{d\hat{t}}\right|_{t=0}=0.\end{aligned}\tag{17}$$

This set of equations and boundary conditions is identical to the form given in (15), and so if a validated solution exists for one problem, it can be used to validate solutions to the other problem.

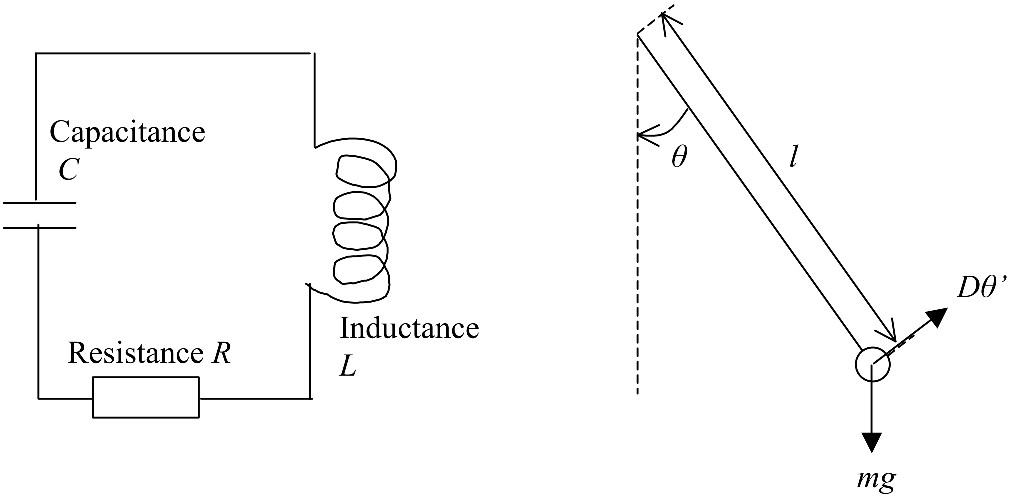

Figure 14: Figures showing (a) the electrical circuit for $t > 0$ and (b) the pendulum, models for which are derived in section 5.2.2.

In order to use this approach of comparing two models with the same non-dimensional form, it is necessary to identify an equivalent problem that has already been solved reliably. As is illustrated by the example above, often the models that are equivalent describe idealised and simplified situations, and when more detailed formulations describing real situations more accurately are introduced the equivalence is destroyed. Thus the technique may be useful as a source of validation data during initial model development, but it becomes less useful as the complexity increases.

5.3 Comparison with other numerical methods

In many cases, more than one numerical method may be suitable for the same problem, but there may be reasons why the user wishes to employ one over another. For example, if users want their solution to the problem to be freely available on the internet, there may be licensing issues if they have used proprietary software. The alternative numerical method can instead be used as a source of validation data. The method usually checks all the details of the solution, since the models are normally designed to provide results at the same space and time points.

There may be other benefits in checking with an alternative calculation method. For example, because the boundary element method is the result of the formulation of a problem as an integral equation rather than a differential equation, the singularities in its results for some problems are of a different type from those generated by finite element software, and so comparison between the two sets of results can be useful for identifying which errors are caused by attempting to approximate a singularity numerically and which are errors that can be removed. There are many applications where comparison of results obtained using finite element, finite difference, and boundary element methods can give a fuller picture of the true behaviour of the system.

As an example, consider the propagation of acoustic waves in an infinite medium. The time-independent velocity potential (a function of position only) obeys Helmholtz's equation

$$\nabla^2\varphi + k^2\varphi = 0,$$

where k is the wave number, $k = 2\pi f/c$, f the frequency, and c the speed of sound.

One problem of interest described by these equations is the calculation of the velocity potential at some location given values of the velocity potential on some closed surface. This problem is particularly useful for the calculation of the "far field" behaviour of a sound source given measurements made in the "near field" (i.e., close to the source). The solution of this problem involves two stages. The first is to calculate the normal derivative of the velocity potential everywhere on the closed surface, using the equation

$$\int_S \frac{\partial G_k}{\partial n_q}(\mathbf{p},\mathbf{q})\varphi(\mathbf{q})dS_q - \frac{1}{2}\varphi(\mathbf{p}) = \int_S G_k(\mathbf{p},\mathbf{q})\frac{\partial \varphi}{\partial n_q}(\mathbf{q})dS_q, \quad \mathbf{p} \in S, \tag{18}$$

where

$$G_k(\mathbf{p},\mathbf{q}) = \frac{1}{4\pi r}e^{ikr},$$

is the Green's function for Helmholtz's equation in three dimensions with $r = |\mathbf{p} - \mathbf{q}|$, S is the closed surface on which values are given, and $\partial\varphi/\partial n$ is the normal derivative of φ with respect to the surface S. This equation is derived by applying Green's second theorem to Helmholtz's equation. Once the values of φ and $\partial\varphi/\partial n$ on S are known, they can be used to calculate values of φ elsewhere in the domain using

$$\varphi(\mathbf{p}) = -\int_S \frac{\partial G_k}{\partial n_q}(\mathbf{p},\mathbf{q})\varphi(\mathbf{q})dS_q - \int_S G_k(\mathbf{p},\mathbf{q})\frac{\partial \varphi}{\partial n_q}(\mathbf{q})dS_q,$$

provided that **p** is a point outside the closed surface.

The problem is ideal for the boundary element method since it is formulated as a surface integral. For certain values of k, this problem becomes singular and has either no solution or an infinite number of solutions. This singularity is only a property of the integral form of the model, and so is not a feature of the physical system. Whilst in theory the problem could be avoided by not solving for these values of k, the quality of the numerical solution obtained by the boundary element method gets significantly worse for values of k that are close to the singular values. Additionally, the problematic values occur more often as k increases, so calculations at high frequency become less and less likely to produce a good solution.

Two methods have been developed to overcome this difficulty. The first is the Schenck method [6] which involves altering the problem by calculating the solution on the surface and at a number of points within the boundary. The solution to this extended problem is calculated in a least-squares sense because the problem is over-determined. The second method is the Burton and Miller method [7], which modifies the problem by taking the derivative of equation (18) with respect to the normal to the surface, multiplying this equation by a coupling coefficient, and adding it to (18).

The Schenck method was validated by checking its results for a problem for which there is an analytic solution over a range of frequencies, and the Burton and Miller method by comparing its results to the Schenck method over the same frequencies.

The differences in the l_2 norm between the analytic method and the Schenck method, and the Burton and Miller method and the Schenck method, are shown in table 2. The differences increase as the frequency increases, due to mesh deficiencies. This decline in quality is shown more strongly when percentage differences are calculated instead of values of the l_2 norm. Results at frequencies above about 6 kHz have an average percentage error of about 15%.

Frequency (kHz)	1	2	3	4	5
Analytic to Schenck	7.37×10^{-4}	1.49×10^{-3}	3.47×10^{-3}	6.54×10^{-3}	9.65×10^{-3}
Schenck to Burton and Miller	3.01×10^{-3}	5.75×10^{-3}	5.36×10^{-3}	3.91×10^{-3}	6.66×10^{-3}
Frequency (kHz)	6	7	8	9	10
Analytic to Schenck	2.15×10^{-2}	3.47×10^{-2}	5.44×10^{-2}	7.89×10^{-2}	7.30×10^{-2}
Schenck to Burton and Miller	1.33×10^{-2}	1.67×10^{-2}	2.51×10^{-2}	7.07×10^{-2}	6.20×10^{-2}

Table 2: Differences in the l_2 norm without averaging, obtained from 481 results, between the analytic solution and the Schenck results, and the Schenck results and the Burton and Miller results. All results are the magnitudes of pressures calculated along a radial line away from a point source.

Figure 15 shows typical percentage differences between the analytic solution and the Schenck method results, and the results of the Burton and Miller method and the Schenck method. These results were generated by modelling a point source and calculating the results along a radial line away from the source. The plot shows the difference in the magnitude of the pressure.

There are several factors that must be considered before using comparison with a different numerical solution as a validation method. One is that the model being used to provide validation data must itself have been validated, at least by using internal consistency checks, but preferably externally as well. Another is that the comparison between the two sets of results should take the accuracy of the validated model into account: there is no point altering the new model until it agrees with the old one to eight decimal places if the original validation of the old model showed it to be accurate only to four figures.

Another drawback is there are cases where the second model may be more accurate than the first, and the advantages of this knowledge are lost through not comparing with an analytic solution. For instance, in table 2, the Burton and Miller method could be significantly better at higher frequencies than the Schenck method if at every point the Schenck method predicted a greater amplitude than the Burton and Miller method, and both were higher than the real value.

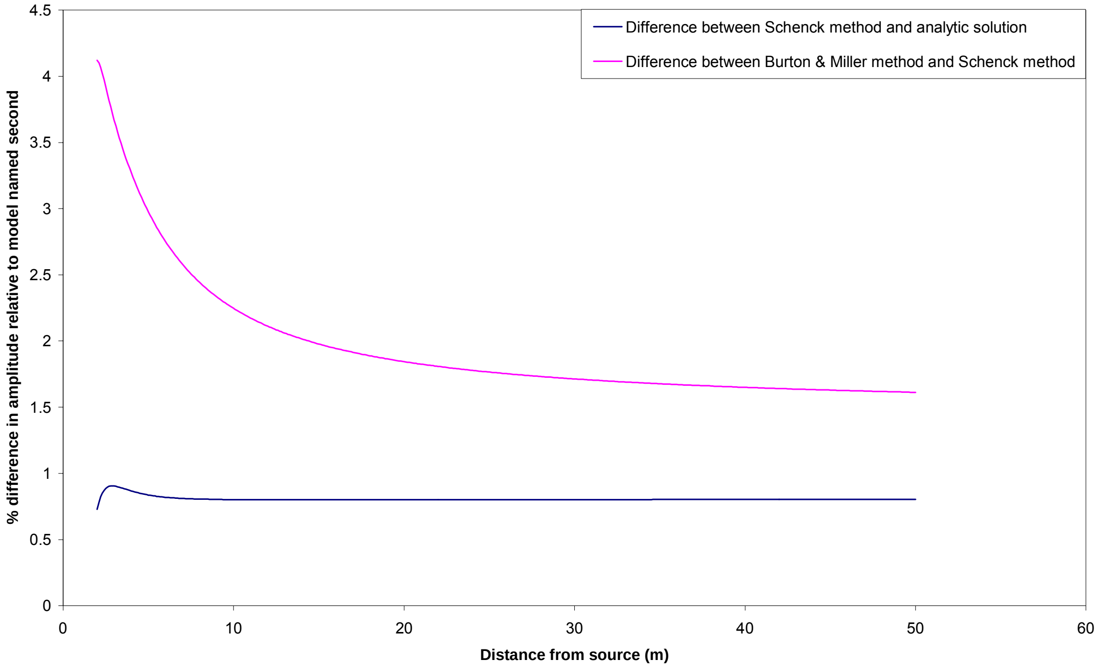

Figure 15: Percentage differences between an analytic solution and the Schenck model, and the Schenck model and the Burton and Miller model for a point source of frequency 2 kHz. Results were generated along a straight line radially away from the source.

6 External consistency checks

External consistency checks are most like the casual definition of model validation, because they check that the model is a reasonable representation of reality. They are the most straightforward to apply, because they generally consist of comparisons between measured data and model results, but they can be misleading if they are carried out without proper consideration of what is required.

6.1 Visual inspection of results

The NPL report "Model validation in the context of metrology" [4] makes the observation that experimentalists often have an intuitive understanding of their measurement data and its errors, and that this insight needs to be incorporated into the validation process, perhaps through formal user requirements documents that identify objectives, assumptions and constraints, and that the growth in certification schemes and quality management systems has accelerated moves towards formal, objective quality requirements rather than expert subjective judgement. These observations mean that if full use is to be made of the metrologist's knowledge of the measurement process, some degree of formalisation of their intuitive checks needs to take place.

The visual inspection of results is one way of including the metrologist's knowledge of the system in the validation process. The method does not produce a quantifiable error. Instead it is a pass/fail criterion, and it usually involves checking global properties of the model rather than results at a specific point. If it is to be regarded as a formal step in the validation procedure, all the visual checks that are carried out need to be documented, so that they can be repeated every time a new model is developed.

"Looks wrong"

Often visual inspection is more a method for invalidation than validation, since it leads to the rejection of results that "look wrong" rather than the complete acceptance of results that "look right". Some of the other methods are extensions of a "looks right" criterion, particularly the conservation checks, since they rely on a metrologist's judgement of the expected behaviour of a system.

The first thing most metrologists would check when making a visual inspection of results is that they are of the right order of magnitude and have the right sign, and that the maxima and minima are in approximately the expected places. Additionally, many results are expected to vary smoothly, if the underlying equations are diffusion-based for example, and often the results are known to be symmetric about some axis which can be checked visually.

For example, consider a cube of perfect uniform material heated uniformly to 100 °C and placed in surroundings where the air temperature is 20 °C, where it cools as a result of radiation and convection. Then it would be expected that the temperature at all points in the cube would lie within the range 20-100 °C, that the cube would be at least as hot as the air at all points, and that the minimum temperature would be at the corners and the maximum temperature in the centre of the cube. Since the heat equation is a diffusion equation, the material is perfect, and no additional heat source is present, it would be expected that the temperature contours would vary smoothly, and that the temperature distribution would be symmetrical. Whilst some of these qualities could be checked in detail using numerical results, they can be checked for visually on a contour plot (see figure 16), as the detailed values of the temperature distribution within the cube is not important for the test. The visual check is used as a brief test that the model is not wrong in any immediately obvious way rather than as a quantitative assessment of model accuracy.

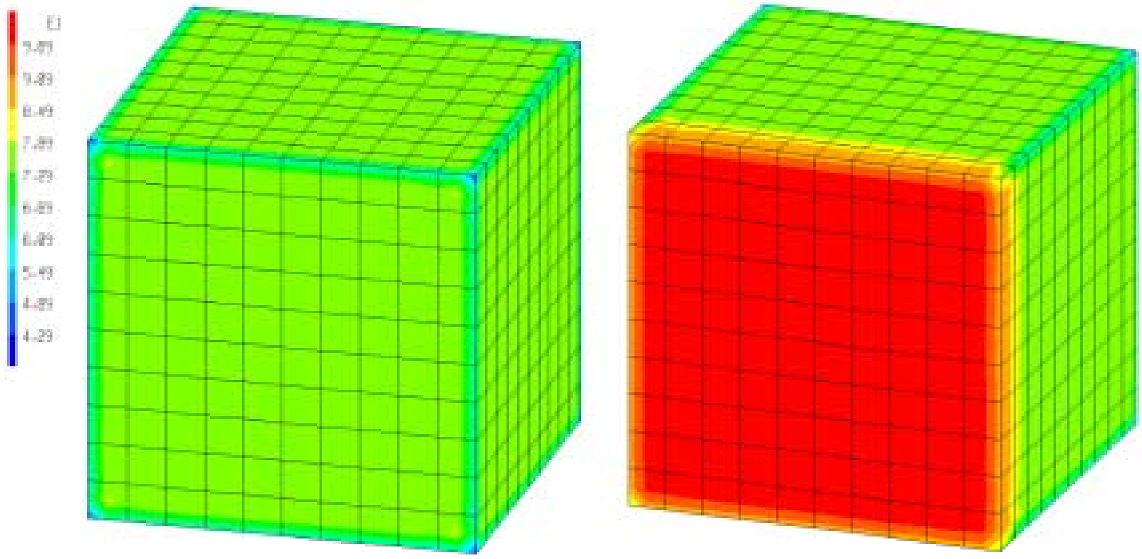

Figure 16: Example of model that “looks right” (left hand cube: results are rotationally symmetric, maximum is internal and minima are at corners) and “looks wrong” (right hand cube: maximum is on the surface, results are not symmetric) for the problem described in section 6.1.

Another way to test using a visual inspection is to look at extreme values of parameters or results at a large distance from the main area of interest. For example, the long-time steady-state behaviour of many dynamic systems is known, and for field problems in electro-magnetics and acoustics there is often a condition that results should tend towards zero at a large distance from any sources. Similarly, many systems have well-understood behaviour for certain values of their input parameters. For instance, many material models can be reduced to simple a elasticity model by appropriate parameter choices.

There may be a non-dimensional parameter that has a well-understood behaviour that can be plotted to give an indication of any modelling errors. For example, in acoustics one might choose *ka*, the product of the wave number and transducer radius, as the variable of interest. This choice of parameter is useful because the solutions of the equations for a pressure field radiated by a source transducer or for the directionality of an acoustic source are frequently functions of the dimensionless variable *ka*, so that any particular solution is valid for all products of *k* and *a* that give the same numerical value; thus the field prediction for $k = 1$ mm^{-1} and $a = 6$ mm should be identical to that for $k = 0.5$ mm^{-1} and $a = 12$ mm. This can be regarded as a visual check in this case because it uses plotted quantities rather than a quantitative analysis of the results.

“Looks right”

Visual inspection can also be used to add credence to partial validations using other methods. For example, often experimental data are only generated for one aspect of a continuous model, and a visual check that the results in the rest of the solution domain “look right” can help to support validation using comparison with experimental data. This approach must be used with caution, since it in no way proves that the results removed from the validated region are correct. Further validation is required to check them.

Attempts have been made during this project to convert the "looks right" criterion into a quantitative measure of a model's validity, but it has generally been found that the difficulty of this task outweighs its usefulness. Generally, visual inspection is useful for dividing models into two classes ("looks right" and "looks wrong"), but it cannot order them within those classes. The main arguments against quantifying the validity of a model using the "looks right" criterion are:

- The technique is essentially subjective: two metrologists may not agree on the "degree of rightness", and may not even agree on the criteria to use, so the technique is insufficiently well-defined to merit numerical evaluation.
- The qualities being sought in the test are generally sufficiently broad (e.g., is the number positive, or is the wave travelling in the right direction) that the results are a positive or negative answer rather than a numerical value. Most tests similar to this criterion that involve numerical comparisons have been listed elsewhere, e.g., conservation laws and comparison with analytic solution.
- Frequently the test involves checking for "looks wrong" and rejecting bad models. In general modellers are not interested in the degree of validity of invalid models, only in why they are invalid. Thus, ranking the rejected models in terms of their invalidity is irrelevant.

However, there are important exceptions to this generalisation, where quantification of "looks right" is possible. In particular, it can be useful to quantify "looks right" when considering model results in the region of singularities. Many common problems include phenomena that lead to infinite values in the mathematical formulation, but not in reality. For example, the application of point loads leads to infinite stresses, and results at angled unfilleted corners are often singular. Frequently, these infinities are understood and accepted by modellers because the results they are interested in occur away from the location of the singularity. However, these infinite results lead to consideration of the effects of the infinite results on results at points removed from the singularity. Generally, the ideal results would be a very localised effect, since often the singularity is caused by an effect at a single point, so it is possible to quantify how "right" the model looks by using the extent of the effects of the singularity as a measure.

For instance, consider the problem shown in figure 17. The right-hand side of the shape is constrained not to move horizontally, the base is constrained not to move vertically, and the top is displaced vertically by a fixed amount. This problem has a stress singularity at the corner marked X. Two plots of σ_{xx} are shown in figure 18, produced by different meshes. The two meshes have almost identical results (plots are to the same scale), and this similarity shows that the less dense mesh can be considered to be producing converged results for most of the model domain.

The effect of the singularity is shown in figure 19, which is a plot of σ_{xx} along the line AB in figure 17 for several different meshes. The lines are labelled with the number of nodes lying along the line AB. From this graph, a numerical value for the extent of effect of the singularity Δs could be estimated, which could be used as a numerical assessment of how "right" the model looks by calculating $\eta = 1/\Delta s$. Then, the better the model "looks", the larger the value of η will be. For instance the mesh with 145 nodes could have a "looks right" value of 0.5, since only the section between $y = 9$ and $y = 11$

seems to have been affected by the singularity, whereas the mesh with 17 nodes "looks wrong" between $y = 5$ and $y = 15$, and so has a "looks right" value of 0.1.

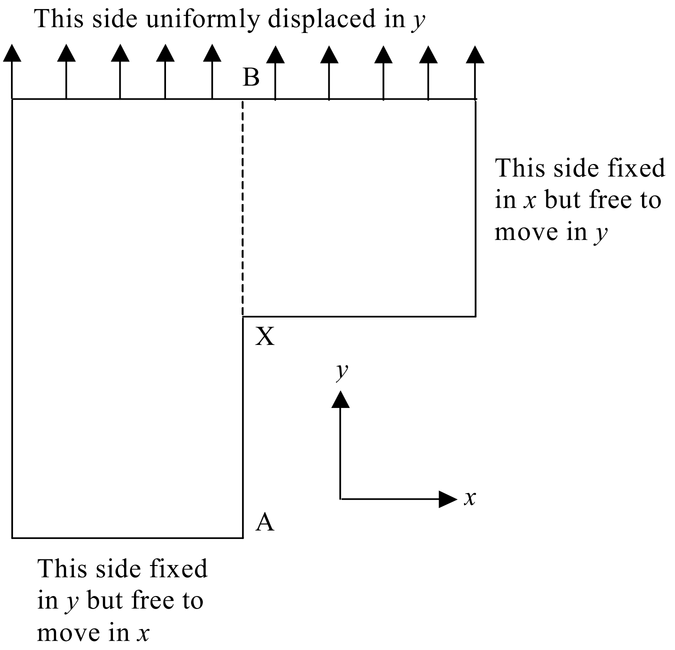
This side uniformly displaced in y
B
This side fixed in x but free to move in y
X
y
x
A
This side fixed in y but free to move in x

Figure 17: Problem including a stress singularity at corner X.

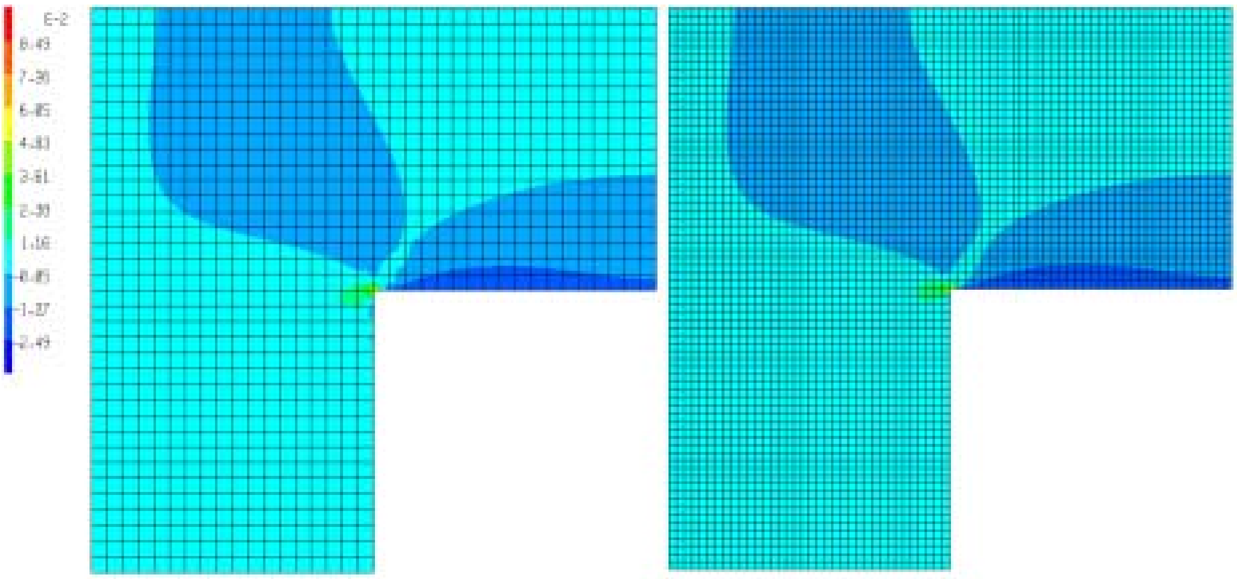

Figure 18: σ_{xx} results from two different meshes (left hand mesh is half the density of the right-hand one). Similarity of the results shows that the mesh density on the left is sufficient to produce converged results in most of the domain.

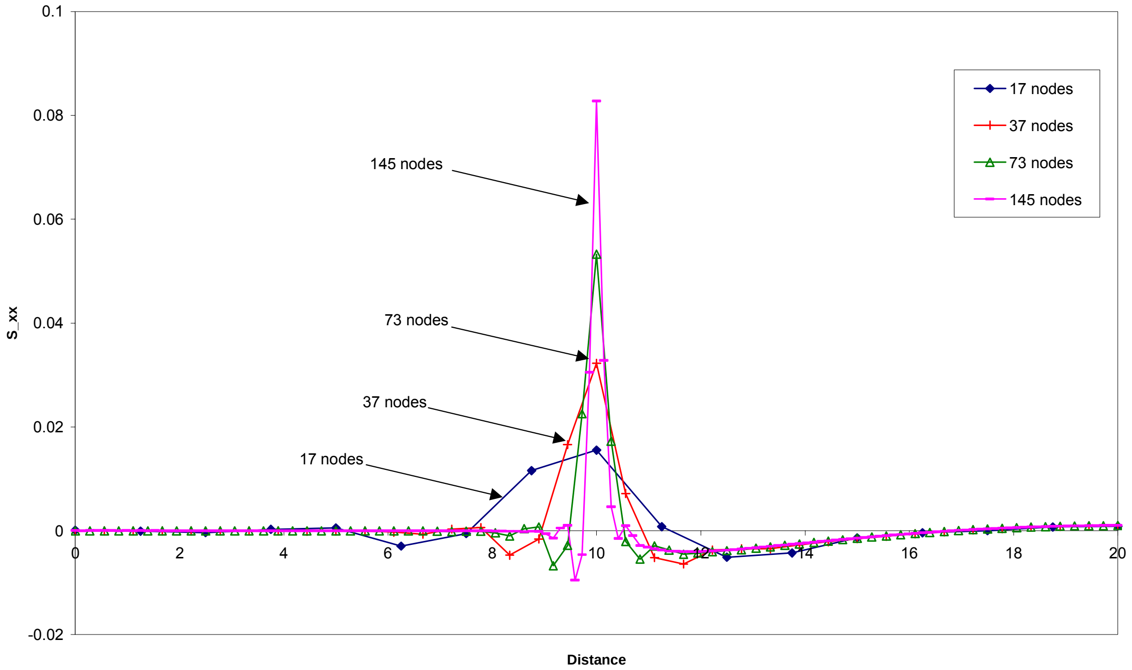

Figure 19: σ_{xx} results along the line AB in figure 17 for different meshes, showing the decreasing extent of the effect of the singularity with increase in mesh density. The left-hand mesh in figure 18 has 73 nodes, and the right-hand one has 145 nodes.

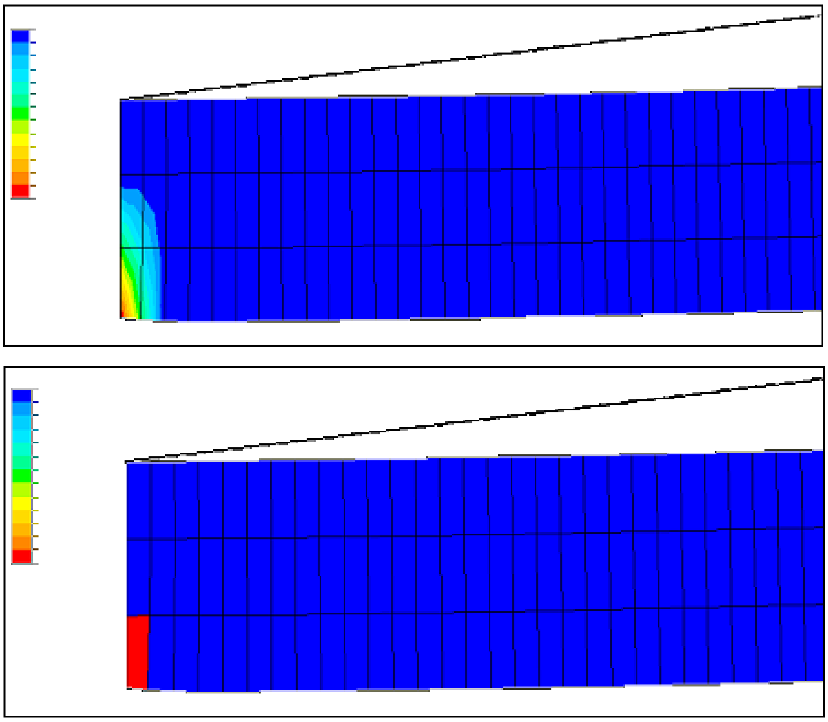

Figure 20a: Contour smoothing activated

Figure 20b: Contour smoothing deactivated

Figure 20: The same results displayed with contour smoothing activated (20a) and deactivated (20b). Results in (20a) are believable to someone with knowledge of the physical process; those in (20b) are not.

The main potential drawback with visual inspection is that care must be taken that the use of post-processing software does not lead to erroneous conclusions. For instance, the two diagrams shown in figure 20 depict the same model at the same time step under the same conditions, but one has had the automatic interpolation of contours option deactivated in the post-processing software. For the quantity of interest, figure 20a is a believable result but figure 20b is not. The best way of examining this data would be to look at the numerical results rather than the contour plots.

Another example is shown in figures 21 and 22. The deformation of a cylinder and piston assembly under stress has been modelled. Prior to the application of the pressure along the gap, the clearance between the two components is 0.32 µm. The model is axisymmetric and the piston radius is 1.25 mm. In figure 21 the piston is at the bottom and the cylinder is at the top. The image purports to show how the piston and cylinder have been displaced. The image indicates that the deformed piston and cylinder overlap.

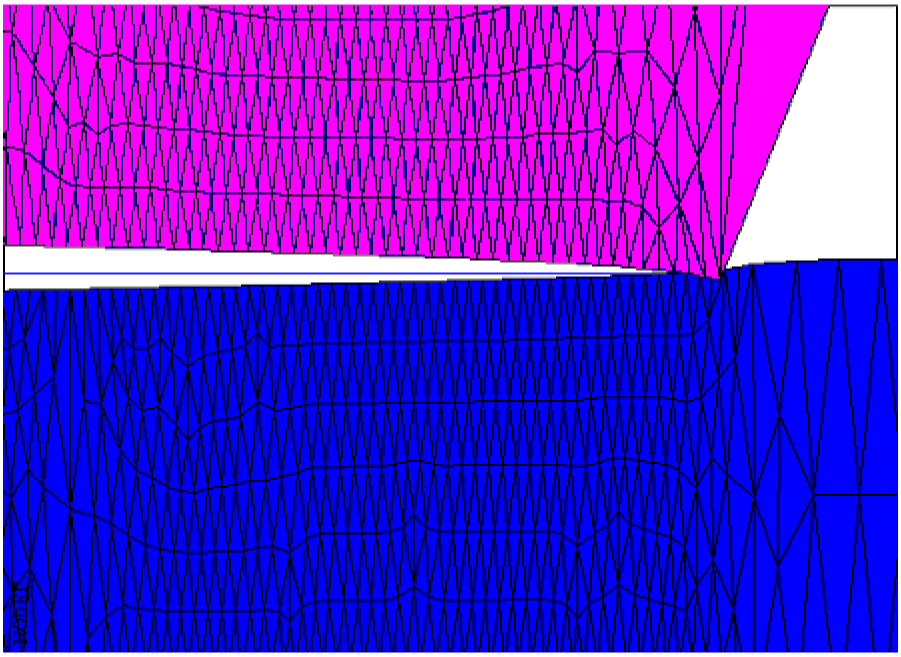

Figure 21: FE post-processing software output showing the piston (blue, below) and cylinder (pink, above) overlapping.

The real gap profile can be obtained by plotting the co-ordinates of the nodes along the gap for the piston and cylinder as shown in figure 22. At the top of the gap the clearance between the two components is 51.5 nm, indicating that the two components do not overlap. This lack of overlap means that the post-processing software output shown in figure 21 plots the displacement results in a misleading way.

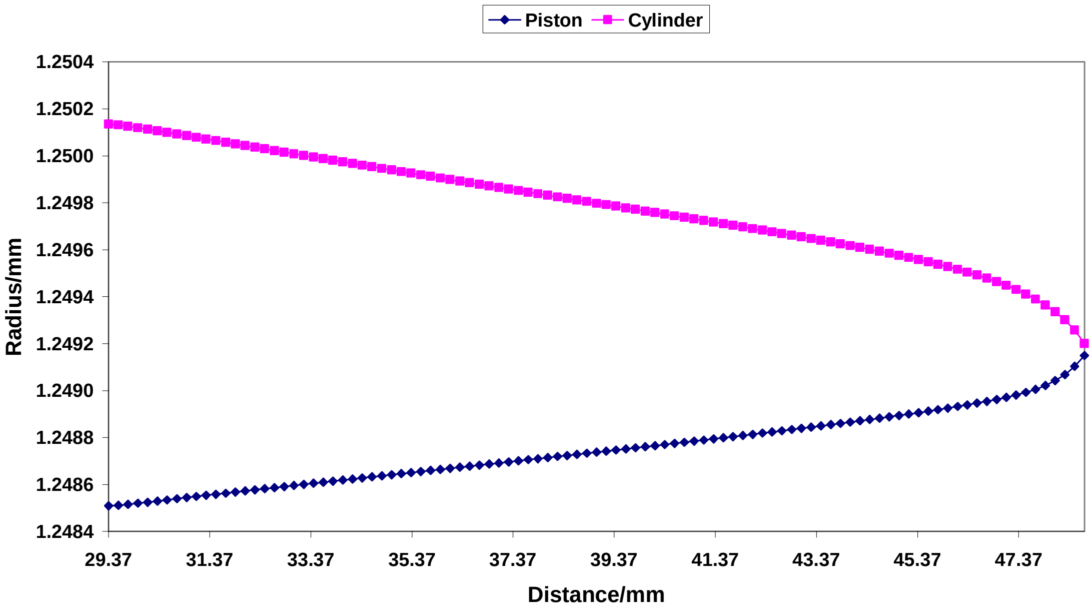

Figure 22: Plot of numerical results obtained from the same FE calculation as figure 21 showing that the piston (blue, lower curve) and the cylinder (pink, upper curve) do not overlap, whereas the post-processing plot shown in figure 21 showed the two components overlapping.

In general, visual output from finite element post-processing software should not be used for drawing conclusions unless the user is certain that the display shows the calculated results correctly, and that all assumptions that went into the creation of the display are fully understood. Understanding the assumptions may only require an understanding of any interpolation carried out by the post-processor, but it is generally preferable to output numerical results and analyse those to obtain conclusions, using the graphics to complement the results.

Another drawback to the visual inspection method is that it relies on the metrologist having the right idea of what "looks right" and what does not. If the metrologist has misunderstood the system a valid model could be rejected unnecessarily. An invalid model accepted in error will probably fail other validation tests, so it is essential to validate using more than just a visual inspection.

6.2 Conservation laws

Several continuous modelling discretisation techniques are derived from conservation laws. The finite volume technique commonly used in computational fluid dynamics software uses the conservation of mass, momentum and energy within a small volume to derive discretised equations. In general, if a quantity is expected to be conserved in reality, then an equivalent discrete version of the same quantity should be conserved by the model. Checks for conservation properties are good for identifying when models are wrong, rather than giving quantified error estimates, and so are essentially pass/fail criteria. They can be thought of as checking for inter-model consistency in cases where the continuous equation is known to conserve some quantity, and external consistency in cases where it is known that the physical system is conservative but less is known about the equations.

Conservation of mass is not an issue for concern for most continuous modelling applications. Many problems do not need mass at all, and some methods (e.g., finite element analysis) applied to problems that do require mass knowledge are structured so that mass is automatically conserved. However, if a method involves a fixed mesh in space with material flowing through it, such as is used for finite volume techniques, mass conservation should be considered.

The main areas where it is important to check conservation of mass are computational fluid dynamics (CFD) and chemical kinetics (particularly in cases where the two are combined, such as combustion problems). Both these areas often involve nonlinear equations and stiff systems (see section 7.1.2 for a definition of stiff in this context) that can require complicated methods to solve and so, even though the continuous equations are often generated from conservation of mass, it is worth checking that the discretised versions preserve the property. Many commercial packages provide output of mass fluxes at each time step, so that the user can check that the solution is behaving correctly.

Conservation of momentum is useful for time-dependent problems such as impact problems. It is also used to derive the Navier-Stokes equations used for incompressible fluid flow. For large models it can be a difficult quantity to calculate, since local velocities within elements may be very different from the net velocity of a body, and hence the summation over all elements may be problematic. Additionally, local velocity estimates can be inaccurate if a body is deforming rapidly, which would skew the results. In general, conservation of momentum may be the least useful of the conservation laws for validation purposes, but the law is sometimes calculated by software packages and so the relevant information may be available to users without any further effort.

Conservation of energy should apply to most models to some degree, although in real experiments there are energy losses that are hard to quantify. It is a good check for any thermal model, although energy lost to the atmosphere in cases where the surroundings are assumed to be at a constant temperature needs careful consideration. It is also useful for stress problems, but applying the law to such models requires monitoring of all strain energies, including elastic, viscoelastic and plastic, as well as any thermal energy.

Many finite element packages enable the evolution of the different types of energy (kinetic energy, recoverable strain energy, non-recoverable strain energy, etc.) to be plotted over the course of a run so that conservation can be checked. Similarly, for a thermal analysis many packages will give heat flux figures so that the balance between heat sources and heat sinks can be checked for a static thermal model to check that the law of conservation of energy is being obeyed.

As an example of checking the conservation of mass, consider the chemical reaction that produces hydrogen bromide,

$$\mathrm{H}_2 + \mathrm{Br}_2 \to 2\mathrm{HBr}.$$

It is clear from this equation that mass must be conserved throughout, since there is the same number of atoms of each element on either side of the arrow. By considering the steps involved in the reaction [21], it can be shown that the differential equations governing the progress of the reaction can be written as

$$\frac{d[\mathrm{H}_2]}{dt} = \frac{d[\mathrm{Br}_2]}{dt} = -\frac{1}{2}\frac{d[\mathrm{HBr}]}{dt} = \frac{k_1[\mathrm{H}_2][\mathrm{Br}_2]^{1/2}}{k_2 + [\mathrm{HBr}]/[\mathrm{Br}_2]}, \tag{19}$$

where [A] denotes the concentration of molecule A, in moles per litre, k_1 and k_2 are constants, and it is assumed that the reaction takes place at constant volume. This equation does not have an analytic solution for general values of k_1 and k_2 , so it requires numerical integration. Since the molecular masses of hydrogen and bromine are known, the total mass can be monitored at all times, since

$$m_{total} = m_{\mathrm{H}_2}[\mathrm{H}_2] + m_{\mathrm{Br}_2}[\mathrm{Br}_2] + m_{\mathrm{HB}r}[\mathrm{HBr}],$$

where m_{HBr} is the molar mass of HBr, etc. The results shown in figure 14 below were generated by two calculations using the same values of k_1 and k_2, but one calculation used the correct equations as in (19) above and the other used

$$\frac{d[\mathrm{H}_2]}{dt} = \frac{d[\mathrm{Br}_2]}{dt} = -2\frac{d[\mathrm{HBr}]}{dt} = \frac{k_1[\mathrm{H}_2][\mathrm{Br}_2]^{1/2}}{k_2 + [\mathrm{HBr}]/[\mathrm{Br}_2]}, \tag{20}$$

so that the rate of change of the concentration of [HBr] is less than that in the correct solution. Accurate values of k_1 and k_2 were not available, so both were arbitrarily set to unity and for this reason no units are given on the graph. It is clear from the graph that the mass is constant in the calculation using (19) but decreases during that using (20).

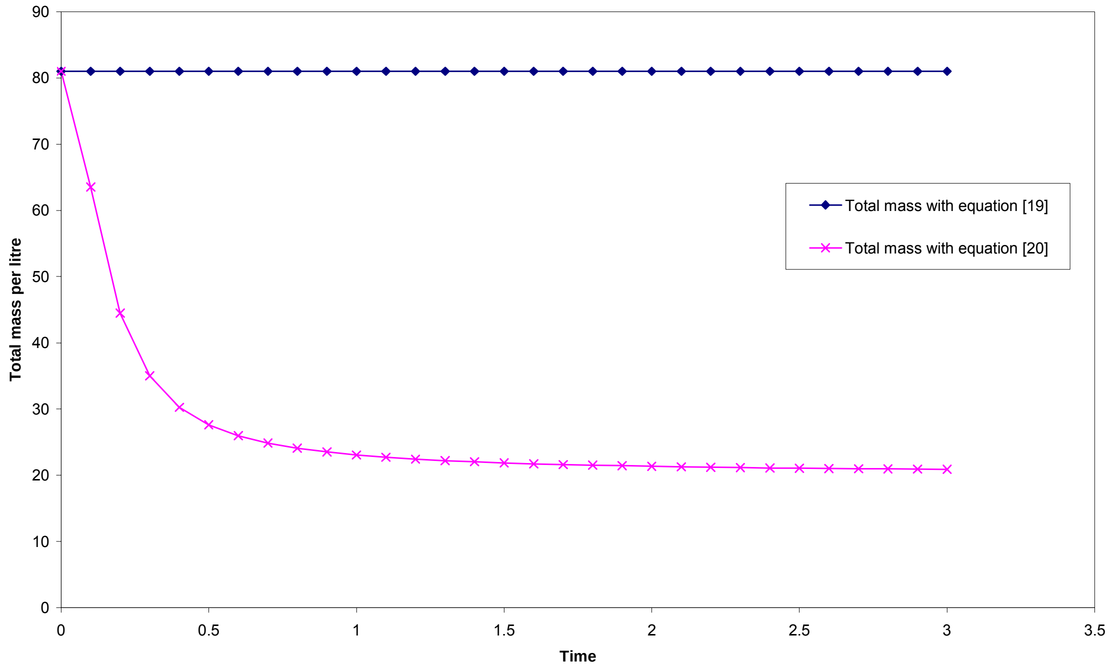

Figure 23: Total atomic mass during reaction to produce hydrogen bromide, calculated using equations (19) (blue line with blobs) and (20) (pink line with crosses). Units are not given due to the arbitrary nature of the constants chosen.

One of the drawbacks of using conservation laws for validation is that they do not examine the details of the results, only an overall property of the entire system. As an example, consider applying the one-dimensional heat equation to a perfectly lagged bar. If it is assumed that no heat escapes from the free ends of the bar, then the total energy should be conserved. The formulation of the problem is

$$\frac{\partial u}{\partial t} = \frac{\lambda}{\rho c_p}\frac{\partial^2 u}{\partial x^2}, \tag{21}$$

$$\begin{aligned}&0 \le x \le 1, \qquad \frac{\partial u}{\partial x}(0,t)=0, \quad \frac{\partial u}{\partial x}(1,t)=0,\\ &u(x,0)=\begin{cases} 2x, & 0 \le x \le \tfrac{1}{2},\\ 2(1-x), & \tfrac{1}{2} \le x \le 1.\end{cases}\end{aligned} \tag{22}$$

If the discretisation method as in equation (5) is used and it is assumed that all material properties are kept constant, it is expected that

$$\sum_{i=1}^{N} U_i^n$$

should be constant for all time steps *n,* as this sum can be regarded as equivalent to energy.

Figure 24 shows the percentage error in the calculated values of this quantity for a stable and an unstable time step. It shows that both time steps conserve this quantity to within 10^{-8}% over the time range considered, so both sets of results would probably be considered to have passed this test. The approximately constant nature of the stable time step errors indicate that these are probably errors due to machine accuracy or software rounding, whereas the increasing errors in the unstable time step are due to algorithm instability.

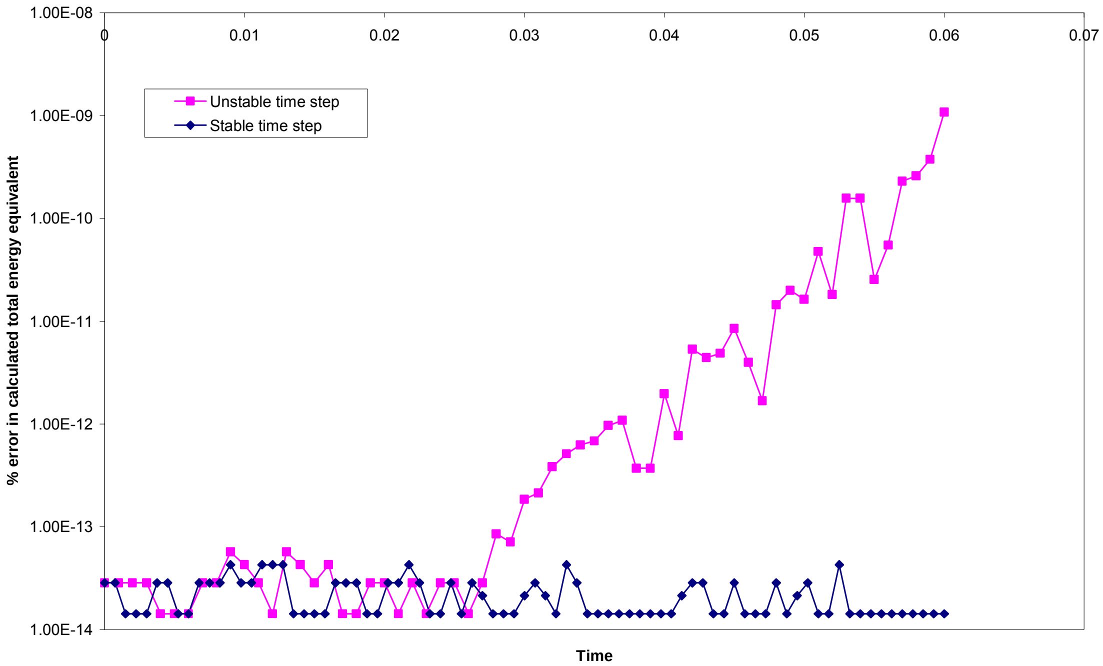

Figure 24: Percentage error in calculated total energy for a stable and unstable time step. The error is plotted on a logarithmic scale, and the maximum value achieved is less than 10^{-8}% over the range considered.

Figure 25 shows the results of both models at *t* = 0.015, and it is clear that, whilst the total energy is conserved, the detailed values are clearly wrong for the unstable time step. This example demonstrates that whilst conservation laws are a useful check, they are not necessarily sufficient to validate a model on their own.

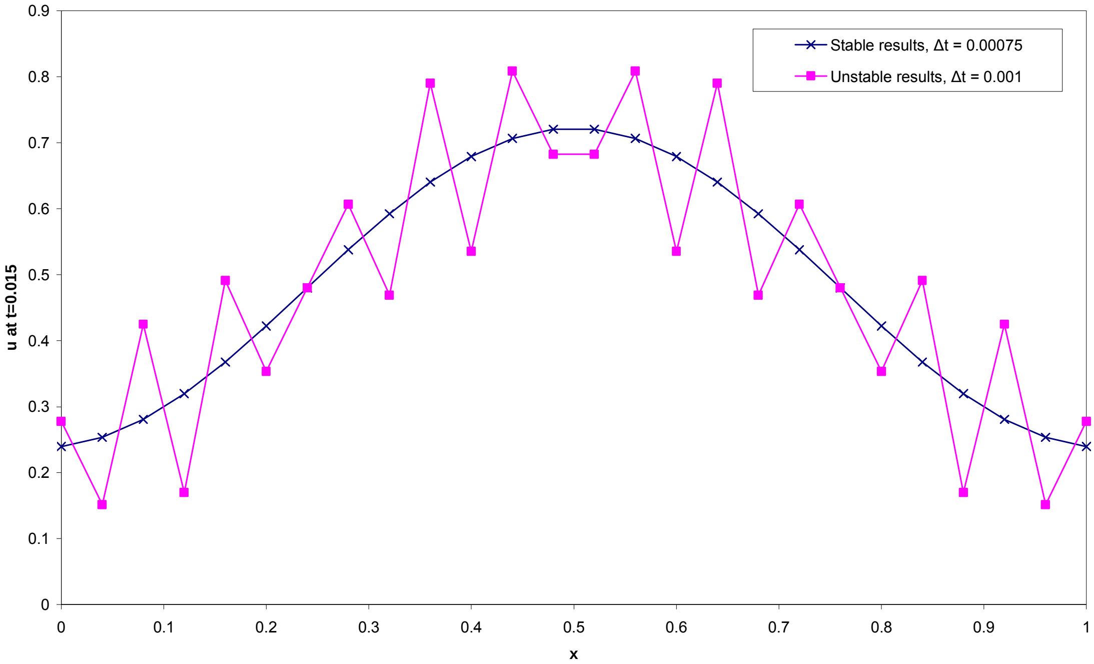

Figure 25: Detailed temperature results at $t = 0.015$, calculated using the discretisation (5) applied to the model (21) and conditions (22)

Another potential problem with use of conservation of energy is that the modeller might inadvertently omit an energy type from the conservation summation and so rejecting the model wrongly. This problem is particularly likely to occur when modelling deformation using complex material models where more than one type of strain energy is involved. Also, many solutions for problems involving non-linearity (e.g., plasticity equations, radiative boundary conditions) are only approximate, so the errors in the solution can lead to apparent energy losses that are due to numerical approximation rather than a modelling error.

6.3 Comparison with experimental data

The comparison of model results with experimental data is probably what the majority of people consider model validation to be. It consists of comparing the model predictions with reality to provide a quantitative measure of how close the model is to the physical situation it is aiming to describe.

For example, consider the experimental set-up shown in figure 26. The hemispherical ram is pushed downwards so that the circular plate deforms, and it can be controlled to move at different speeds so that deformation data at different strain rates is produced. There is a force transducer on the upper ram that enables the force required to deform the plate to be measured over time. Figure 27 shows the comparison between experiment and the results of two finite element models using different models for plasticity over the first 10 mm of deflection for a 1 ms^{-1} test. As well as the plot, values for the root mean square absolute difference between the predictions was calculated, and gave 0.08 kN for the linear Drucker-Prager model and 0.06 kN for the von Mises model.

This plot of the data illustrates some of the problems that are often encountered when trying to validate against experimental measurements. The first problem is that the data are very noisy due to the ram moving at high velocity. The data are actually more noisy than is shown: the plotted line has had an average taken across all measured force values for a fixed deflection. A more suitable method of dealing with these data may have been to attempt some form of signal processing to remove the noise. This method is discussed in more detail below.

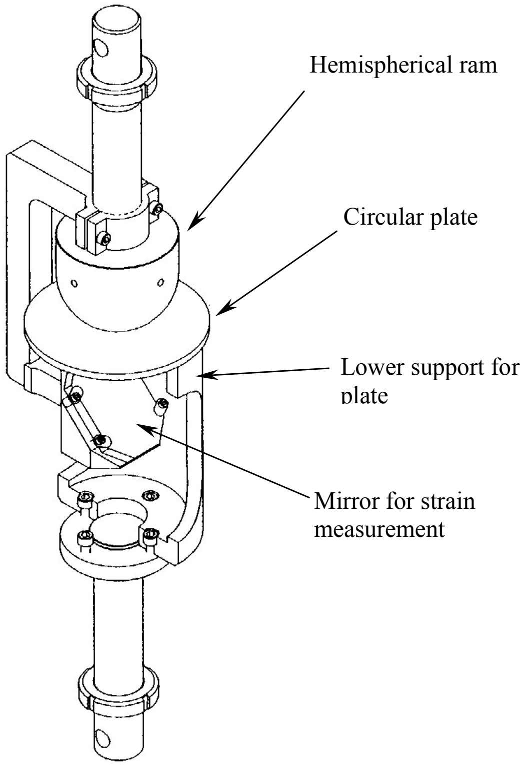

Figure 26: Experimental set-up for force and strain measurement of a centrally loaded plate.

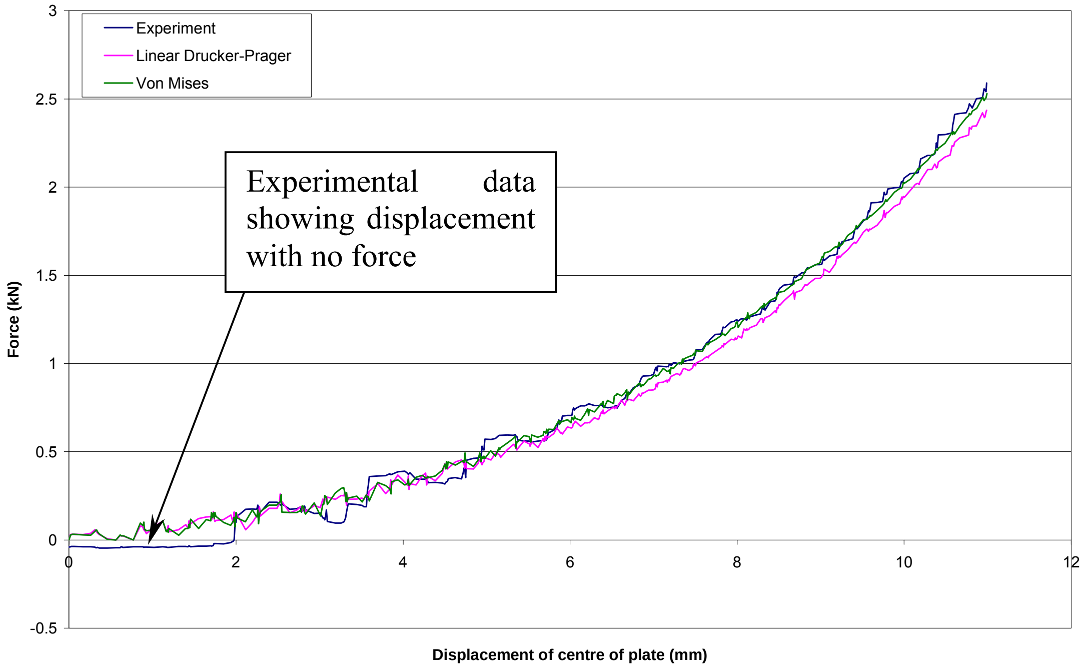

Figure 27: Comparison of force-deflection measurements and model predictions at 1 ms^{-1} for the linear Drucker-Prager and von Mises plasticity models.

Another problem with this data that is mentioned below is the problem of ensuring that the experimental data and model results are synchronised correctly. The force and displacement data were gathered over time and then plotted against one another, and the data capturing devices were not synchronised. Additionally, the data capturing devices were activated before the ram was moved, so that data capture was in progress before the event being modelled had begun. The data shown above contain a small straight line below the displacement axis which indicates displacement with no force. This feature of the data is probably due to imperfect initial contact between the plate and the ram, but it makes it difficult to be sure that the correct time zero has been chosen for the force data.

Continuous models produce results over their entire domain of calculation. Often measurement data only exist for a small area of this domain. When comparing results with experiment, careful consideration needs to be given to whether the results being compared are sufficient to validate the model over the entire domain. It may be that this degree of validation is not needed, but it is possible that the results in other areas may affect the results in the area of interest under different experimental conditions. It is useful in such cases to have a statement of the areas of the domain for which the model has been validated.

Limits of comparison with experimental data

Some continuous models can calculate more than one type of result, for instance finite element stress analysis calculates displacements, stresses and strains, and thermal analyses calculate heat fluxes as well as temperature distributions. Validation of one type of result does not necessarily mean that the other types are validated as well, particularly if they are linked in a complicated or non-linear manner.

As well as the force transducer on the upper ram, there is a mirror beneath the lower surface of the plate making it possible to calculate the strain there, giving two separate sets of validation data. Figure 28 shows the comparison between experiment and one of the finite element models for the strain data. From this figure and figure 27 it is clear that whilst the force data is modelled well, there are significant discrepancies between the strain data and the model results. The point of these figures is to show that use of a global property of the model for validation does not guarantee validation of the details of the model. A similar point was made in section 6.2 on use of conservation laws for validation. The force is a global property of the model and using it to validate the whole model is similar to using a conservation law. These figures show that it is possible for the global properties of two models to be identical, but their local details to differ significantly.

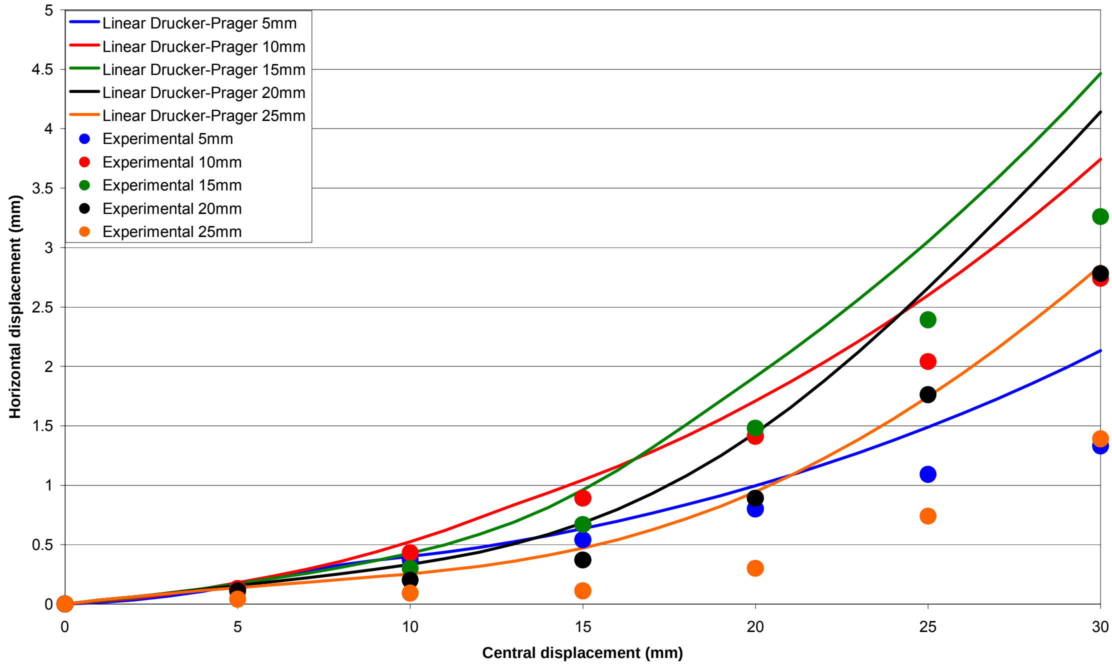

Figure 28: Comparison of strain measurements and predictions for the slowest test speed at different radii on the plate.

Sometimes it can be worth considering designing a simplified form of the experiment that will obey as many of the assumptions made in the model as possible. Doing so will result in a set of data that may be of little use of itself to the metrologist, but that will provide good validation data for the model. It will give a good picture of the best agreement possible between model and experiment, so that when comparison with the full experiment is carried out, it should be possible to investigate how significant the simplifying assumptions are. This technique can be very useful when used to simplify geometries, boundary conditions, and time-dependent problems. Even where the development of a simplified model is not possible, it is useful for the metrologist and the modeller to try to estimate, at least to an order of magnitude, the likely size of the effects that they have ignored when developing the model by identifying an expression that describes the neglected terms and evaluating it using any approximations necessary. For instance, the likely effects of neglecting radiative heat loss can be evaluated by estimating the likely maximum and ambient temperatures and applying the Stefan-Boltzmann law, and the size of this term can be compared to the other heat flux terms used in the model.

It is inevitable that model results and experimental data will not be identical. There are a number of reasons for this, including:

- measurement uncertainty
- model uncertainty and error
- uncertainty about "time = 0"

There are further problems produced when attempting to take these factors into account. A key point is how likely it is that the error and uncertainty estimates are accurate. Some error bounds produced by methods described in this report produce over-estimates of the actual errors, which may result in false acceptance of a model.

Similarly an over-estimate of the measurement uncertainty could lead to false acceptance of model results.

Factors affecting use of experimental data for validation

There may be some doubt as to when exactly the experiment started: often measurement devices start to record data before the event the experiment is aimed at capturing takes place. Conversely, there may be a delay between the occurrence of the event and the first measured response to it. This delay produces an additional measurement uncertainty on the time axis, which needs to be considered.

Noisy data can be dealt with in several different ways. The first is to apply some form of smoothing technique, such as digital signal processing methods, filtering, curve fitting, or local averaging. Whilst these techniques will produce something that can more easily be compared with model results, they impose further assumptions on the data. The modeller and metrologist need to be sure that these assumptions are reasonable for the system in question. The second way is to choose carefully the error norm applied to the results. Error norms are discussed more fully in section 2.2, but it could be that the validation decision needs to be based on more than one norm. Another way is to look only at the averaged or integrated properties of the system. Doing so is similar to looking at conservation laws, in as much as two models can have identical averaged properties but different details, which may lead to false acceptance. If, however, the primary result required from the model is an averaged quantity, a model that has been so validated may be acceptable. A more complicated way is to incorporate some form of noise term in the model itself. If the structure of the noise is reasonably well-understood this can lead to improved comparisons. In general, the metrologist and modeller need to consider what the sources of noise are likely to be, and see which of these options is likely to lead to an improvement in comparisons without imposing further assumptions on the data.

It may be possible to improve the comparability of prediction and experiment by only using a subset of the collected data. For example, some models and experiments will not agree near boundaries, corners and interfaces, but results can be satisfactory away from such regions. Also, for data gathered over time there may be transient effects that are neglected in the model that will adversely affect the results initially, but will decay as time progresses, so that later results will be a better choice for comparing to a model. This judicious choice of results for use in the validation process is sometimes simpler than trying to model short-term transient effects, and often data gathered after a longer time is less noisy. However, this choice does mean that any short-term effects that the model is intended to include are not being validated, so the choice should only be taken when no important effect is likely to be ignored.

Choice of validation data

There is a danger of circularity of argument when validating against experimental measurement. If the same data are used for determining model parameters and validating the model, whilst it may be possible to assess the overall suitability of the chosen model, it is not possible to say anything about the values of the parameters other than that they have been determined correctly initially.

Many mathematical models in metrology are used to assist with uncertainty estimation. In particular, models are used to look at Type B evaluations of uncertainty and to estimate correction factors that are applied during data processing. Often in such cases the model is used to describe effects that it has not been possible to quantify experimentally. This usage of the model means that it is difficult to be sure that the model results describe the effect correctly if only comparison with experiment is used for validation, since some other unidentified phenomenon could be present. The model may identify effects that need further experimental investigation, but it is inadvisable to validate a description of a phenomenon that has not been quantified experimentally against measured data.

Another problem with validation against experiment is that often some model parameters cannot be obtained for the experimental set-up. In particular, material properties for the internal parts of equipment are often very hard to obtain. Frequently in this situation values from reference books are used as best estimates, but doing so causes two further problems. The first problem is that most book values do not have uncertainties associated with the values given, and the second is that often properties for the same material vary widely according to its exact structure. For example, the properties of steel strongly depend on the chemical composition and the method of manufacture. Both these problems tend to lead to a disregard of the effects of material property uncertainty. They can also lead to material properties being “adjusted” until data and prediction match, which is not advisable unless more than one data set requires the same adjustment as the modified parameters may be hiding a different fundamental flaw in the model.

7 Inter-model consistency checks

Inter-model consistency checks are in some ways very similar to internal consistency checks because they check the validity of an implementation of the model. As was explained in section 1.1, the methods check that the chosen discretisation of the continuous problem is valid. Since the methods are concerned with discretisations, they usually take the form of tests of the mesh and its properties. The modeller generally has a large amount of control over the form of the mesh, particularly over the typical element size, and so inter-model consistency checks can be of great benefit because they can give guidance on how to alter the mesh to reduce the discretisation error.

Inter-model consistency checks are generally quite mathematically involved since they are based on using the important properties of the algorithms and methods to which they are applied. Some software packages implement them to provide adaptive meshing and error estimates for model solutions.

7.1 A priori methods

A priori methods aim to guarantee that the approximate solution obtained by a method will converge to the exact solution as the space (and, if necessary, time) step sizes tend to zero. Such convergence is vitally important, since a model that does not converge to the correct solution of the problem it is trying to solve is generally pointless. The methods validate local properties of models and often provide global error bounds as well.

A priori methods often make assumptions about the existence and behaviour of the exact solution to the problem. Since in general nothing detailed is known about the exact solution, a priori methods can produce information about general trends for solutions but do not usually produce quantifiable error estimates. They usually prove that the differences between an approximate numerical solution and the exact solution for a problem at the mesh points are bounded above by some positive power of the typical mesh dimension.

Generally, a priori methods are based on determining stability and consistency properties of models (see section 2.3 for definition of these terms). The Lax Equivalence Theorem [3, p139] states that for a linear well-posed problem and a linear approximation that is consistent with it, the stability of the scheme is necessary and sufficient for its convergence. This means that proofs of consistency and stability are enough to prove convergence for linear problems. Whilst the theorem does not hold for non-linear problems, consistency and stability are still valuable properties for an approximation to have, since consistency means that the discretisation error will tend to zero as the mesh size decreases and stability means that the round-off errors caused by finite-precision arithmetic will not grow to swamp the true solution.

Often it is easier to calculate an upper bound on $\left\| \mathbf{U}^n - \mathbf{u}^n \right\|$ than it is to calculate $\mathbf{T}^{\mathrm{n}}$ and a stability criterion directly, particularly when the method being studied is not a finite difference approximation. For instance, there is a large amount of a priori work [27] that has calculated bounds on the errors of finite element approximations of various types in various norms by considering $\left\| \mathbf{U}^n - \mathbf{u}^n \right\|$ directly. Such work usually requires rigorous numerical analysis and careful consideration of the assumptions being made about $\mathbf{u}$ at each stage of the argument, particularly for problems in more than one dimension, and so may not be of interest to many modellers. However, having a bound on the error in terms of mesh size parameters means that a model can be validated by reducing the mesh size and checking that any changes in the solution are insignificant. If this is the case, the mesh can be regarded as being sufficiently dense to give a converged solution for that problem. These are mesh convergence tests, as described in section 7.1.3 below.

It should also be noted that convergence is not sufficient to guarantee good results. For example, consider the equation

$$\begin{aligned}&\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}=0, \qquad x>0,\\ &u(0,t)=0, \\ &u(x,0)=u^0(x)=\exp\left(-250\{x-0.25\}^2\right),\end{aligned} \tag{23}$$

where a is a positive constant. This equation is useful in transport problems where diffusion is less significant than advection. This problem has the exact solution

$$u(x,t)=u^0(x-at), \tag{24}$$

which is a translation of the initial conditions along the x-axis at a constant speed. Consider applying the upwind method to it:

$$U_j^{n+1}=U_j^n-a\frac{\Delta t}{\Delta x}\left(U_j^n-U_{j-1}^n\right). \tag{25}$$

Linear stability analysis shows that this method is stable for $a\Delta t \le \Delta x$ and has a truncation error

$$T_j^n=\frac{1}{2}\left(\Delta t\frac{\partial^2 u}{\partial t^2}-a\Delta x\frac{\partial^2 u}{\partial x^2}\right)+O\left(\Delta t^2,\Delta x^2\right), \tag{26}$$

which will tend to zero as Δt, Δx, tend to zero (full details are available [3, chapter 4]). It can also be shown that the method conserves the energy measure

$$\sum_{j=1}^{M}\Delta x U_j^n ,$$

so a check for conservation of energy would be passed. The case a = 0.7 has been considered with $\Delta t = \Delta x$. The results are as shown in figure 29 for t = 0.64, calculated using three different time steps. As can be seen, even for the smallest time step the results are not good. The poor quality of the results is due to numerical dissipation which causes spreading out of the wave. Since the truncation error is of the form (26), which can be rewritten as

$$T_j^n=\frac{1}{2}\left(a^2\Delta x\frac{\partial^2 u}{\partial x^2}-a\Delta x\frac{\partial^2 u}{\partial x^2}\right)+O\left(\Delta t^2,\Delta x^2\right),$$

for a a positive constant and Δt=Δx (from repeated applications of the original PDE), the upwind scheme can be considered to be solving a problem of the form

$$\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}=\frac{-a(1-a)\Delta x}{2}\frac{\partial^2 u}{\partial x^2}, \qquad x>0,$$

and so if the coefficient $a(1-a)\Delta x/2$ is not sufficiently small, the dissipative term will be significant and results will be like those shown in figure 29.

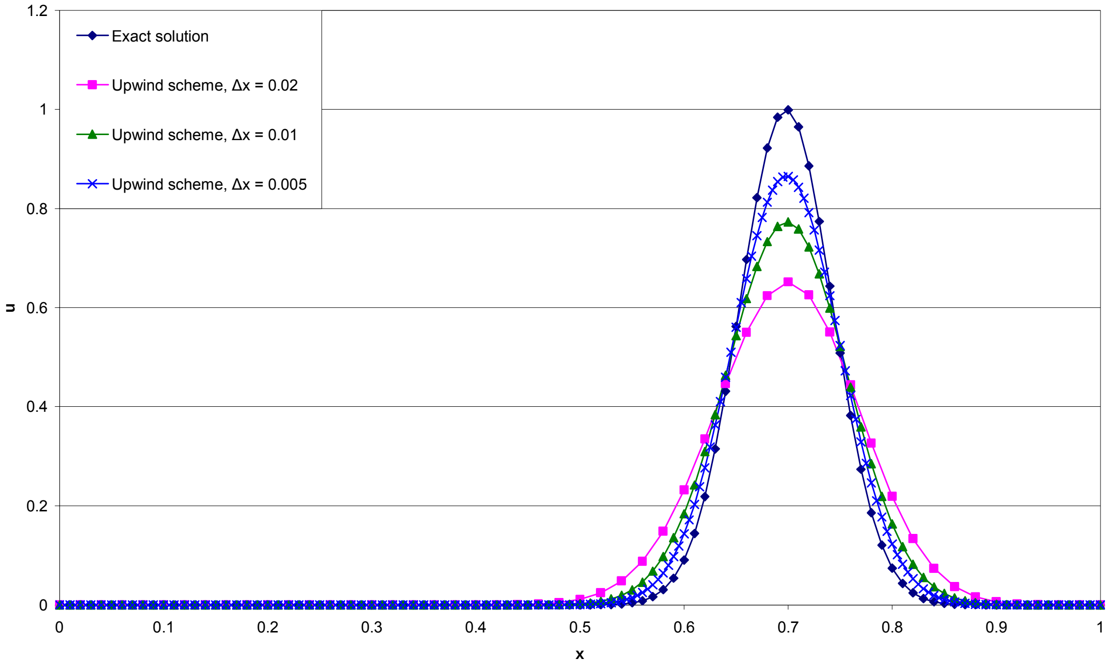

Figure 29: Exact solution to (23) for a = 0.7, and three different approximations calculated using the upwind scheme with $\Delta t = \Delta x$. Results are shown for $t = 0.64$.

7.1.1 Consistency

As was explained in section 2.3, consistency ensures that the truncation error $\mathbf{T}^n$ generated over a single time step will tend towards zero as the mesh size decreases. Thus, an improved solution can be obtained by decreasing the mesh size (provided that numerical rounding errors do not become significant), and proving consistency is a key step in proving convergence, an important property for an approximation method.

Often, $\mathbf{T}^n$ is calculated by expanding all terms in each line of the method (2) as Taylor expansions about some central point, reducing the expansion to its simplest for, and attempting to find upper bounds on the mesh-independent parts of the resulting expression. Most finite difference methods have been developed by using Taylor series expansions to determine expressions for derivatives, so this is often straightforward for such techniques.

Proof of consistency is often considerably less straightforward for other approximation techniques, as they deal with more arbitrary meshes, but often other methods can be used to bound the truncation error in some norm for other problems. In particular much work has been carried out on proving the consistency of the finite element method using maximum principles in the energy norm, and most finite element theory books (for example Zienkewicz and Taylor [8]) give some details of truncation errors of different element formulations as well as proofs of the general properties of solutions.

As an example of a proof, consider the problem

$$\begin{aligned} &\nabla^2 u(x,y) = f(x,y), \quad 0 \le x, y \le 1, \\ &u(0,y) = u(1,y) = u(x,0) = u(x,1) = 0, \end{aligned} \tag{27}$$

and apply a five-point finite difference approximation method to it so that

$$\frac{U_{i+1,j}+U_{i-1,j}-2U_{i,j}}{\Delta x^2}+\frac{U_{i,j+1}+U_{i,j-1}-2U_{i,j}}{\Delta y^2}=f_{i,j}, \qquad 1\le i\le M, 1\le j\le N, \tag{28}$$

where $f_{i,j} = f(i\Delta x,\ j\Delta y)$. Further details of the analysis underlying the following are available [3, chapter 6].

Assuming that all the input parameters are sufficiently smooth in x and y that u is at least four times differentiable, the solution in the region of $u(i\Delta x, j\Delta y)$ can be expanded as a Taylor series, so that

$$u(x+\varepsilon,y)\approx u(x,y)+\varepsilon\left.\frac{\partial u}{\partial x}\right|_{x,y}+\frac{\varepsilon^2}{2!}\left.\frac{\partial^2 u}{\partial x^2}\right|_{x,y}+\frac{\varepsilon^3}{3!}\left.\frac{\partial^3 u}{\partial x^3}\right|_{x,y}+\frac{\varepsilon^4}{4!}\left.\frac{\partial^4 u}{\partial x^4}\right|_{x,y}+O(\varepsilon^5), \tag{29}$$

and a similar expansion exists for $u(x, y + \delta)$. Then the truncation error as defined in section 2.3 is

$$\begin{aligned}T_{i,j}&=\frac{u_{i+1,j}+u_{i-1,j}-2u_{i,j}}{\Delta x^2}+\frac{u_{i,j+1}+u_{i,j-1}-2u_{i,j}}{\Delta y^2},\\&=-\left(\frac{\Delta x^2}{12}\left.\frac{\partial^4 u}{\partial x^4}\right|_{\substack{x=x_i\\y=y_j}}+\frac{\Delta y^2}{12}\left.\frac{\partial^4 u}{\partial y^4}\right|_{\substack{x=x_i\\y=y_j}}+O(\Delta x^4,\Delta y^4)\right), \quad 1\le i\le M, 1\le j\le N.\end{aligned} \tag{30}$$

The truncation error is a space discretisation error caused by truncating the infinite Taylor series after six terms. Let the maximum value of $|T_{i,j}|$ over $1\le i\le M,\ 1\le j\le N$ be T. Then by the mean value theorem applied to T, there must be two points (ξ_1, η_1) and (ξ_2, η_2) in the domain $0\le x\le 1,\ \ 0\le y\le 1$ such that

$$T=\frac{\Delta x^2}{12}\left.\frac{\partial^4 u}{\partial x^4}\right|_{\substack{x=\xi_1\\y=\eta_1}}+\frac{\Delta y^2}{12}\left.\frac{\partial^4 u}{\partial y^4}\right|_{\substack{x=\xi_2\\y=\eta_2}}.$$

Since $T = C_1\Delta x^2 + C_2\Delta y^2$, $T\to 0$ as Δx, $\Delta y\ \to 0$, the method (28) is consistent with the problem (25). The problem is static and consistency has been shown, so, from the Lax Equivalence Theorem, the method is convergent for the problem (27). The value of T is unknown, and since it depends on the form of the exact solution an upper bound on the error is not known explicitly, but the knowledge of the form of dependence of T on Δx and Δy is useful for mesh convergence tests (section 7.1.3) and Richardson extrapolation (section 7.2.3).

7.1.2 Stability

Time step stability analysis aims to ensure that unstable time steps are not used for transient models. If an unstable time step is chosen, round-off errors will grow exponentially and often rapidly swamp the true solution so that nothing of value is gained from the model. This unlimited growth of errors is particularly problematic since useless results are often only discovered after many hours of CPU time when the final solution is checked.

Many finite element software offer the user a choice of implicit or explicit solution methods. Explicit methods are those in which the solution at the next time step is determined only from the calculated values of the solution at the previous time steps. In equation (2), these are methods in which **A** is the identity matrix. These methods are usually easy to program and can solve individual time steps rapidly, since all they require is matrix multiplication and vector addition, but sometimes they have tight restrictions on the time step and so calculations require a large number of time steps, which can lead to long run times. Implicit methods are those in which **A** in (2) is different from the identity matrix, and so the solution of such problems requires matrix inversion, which can be computationally expensive for large **A**. However, implicit methods generally have a much less severe restriction on the time step, and so they need fewer time steps to reach a solution.

Automatic time step generation

Most finite element packages that perform transient analyses estimate a stable time step automatically, and many update this calculation during the analysis so that the model runs efficiently (this is called adaptive time stepping and is described more fully in section 7.2.2). The automatic time step can often be overwritten by the user, but this option should not be taken unless the time step is to be decreased so that results can be saved more frequently. Unstable time steps invariably lead to poor results, but sometimes the errors can take a long time to accumulate, so that much time and computational effort is wasted before the user realises the results are worthless. The automatically generated time steps are usually on the conservative side, and this conservatism combined with the updating during the run mean that the automatic choice is usually safer and more efficient.

The automatic time steps in the FE software package ABAQUS Explicit [16] are generated based on the fact that for a linear problem, $\Delta t \leq 2/\omega_{\max}$ will be a stable time step, where $\omega_{\max}$ is the highest frequency of the system. Initially this value is calculated by estimating the highest frequency of each element in the model, and the largest of these frequencies is always an upper bound for the highest frequency of the entire system. This means that the initial estimate of the time step will be an over-estimate, which could lead to excessive run times if that estimate were to be used throughout. During the model run, a more accurate estimate for the entire model's highest frequency is calculated so as to reduce the number of steps needed.

For simpler models, an upper bound for the stable time step can be calculated by hand. Consider seeking a solution of equation (3) (ignoring boundary conditions at present) that has the form $u(x,t) = f(x)g(t)$. Separating the two parts of the solution and solving for each shows that any such solution will be of the form

$$u(x,t) = \{a_k \sin(kx) + b_k \cos(kx)\}e^{-k^2\alpha t}, \qquad (31)$$

for some real constant k, where $\alpha = \lambda/(\rho c_p)$. Writing the initial conditions as a Fourier series,

$$u(x,0) = \sum_{m=1}^{\infty} a_m e^{im\pi x}, \qquad (32)$$

it is clear that $k = m\pi$, and so

$$u(x,t) = \sum_{m=1}^{\infty} a_m e^{-m^2\pi^2\alpha t} e^{im\pi x} \qquad (33)$$

is a solution to (3).

To seek a solution to the discretised problem (5) that is of a similar form, try

$$V_j^n = \sum_{m=0}^{\infty} a_m \lambda^n e^{im\pi(j\Delta x)} ,$$

where λ represents the time evolution and may be a function of m. Substituting this expression into (5) gives

$$\begin{gathered}\sum_{m=0}^{\infty} a_m \lambda^{n+1} e^{im\pi(j\Delta x)} = \\ \sum_{m=0}^{\infty} \left\{ a_m \lambda^n e^{im\pi(j\Delta x)} \left(1 - 2\frac{\alpha\Delta t}{\Delta x^2}\right) + \frac{\alpha\Delta t}{\Delta x^2} \left(a_m \lambda^n e^{im\pi(\{j+1\}\Delta x)} + a_m \lambda^n e^{im\pi(\{j-1\}\Delta x)}\right)\right\}.\end{gathered} \tag{34}$$

Since the terms $e^{im\pi j\Delta x}$ are linearly independent, the terms in the summations on either side of (34) can be considered individually. Dividing a single term throughout by $a_m\lambda^n e^{im\pi j\Delta x}$ gives

$$\begin{aligned}\lambda(m) &= \left(1 - 2\frac{\alpha\Delta t}{\Delta x^2}\right) + \frac{\alpha\Delta t}{\Delta x^2}\left(e^{im\pi(\Delta x)} + e^{-im\pi(\Delta x)}\right) = 1 + 2\frac{\alpha\Delta t}{\Delta x^2}(\cos\{m\pi\Delta x\} - 1), \\ &= 1 - 4\frac{\alpha\Delta t}{\Delta x^2}\sin^2\left\{\frac{m\pi\Delta x}{2}\right\}.\end{aligned} \tag{35}$$

Hence if there are two solutions V and W corresponding to initial conditions $V^0(x) = \sum_{m=1}^{\infty} a_m e^{im\pi x}$ and $W^0(x) = \sum_{m=1}^{\infty} b_m e^{im\pi x}$, then

$$V_j^n - W_j^n = \sum_{m=0}^{\infty} (a_m - b_m)\lambda^n e^{im\pi(j\Delta x)},$$

and the same expression for λ is used for both. Hence

$$\begin{aligned}\left|V_j^n - W_j^n\right| &= \left|\sum_{m=0}^{\infty} (a_m - b_m)\lambda^n e^{im\pi(j\Delta x)}\right|, \\ &\leq \sum_{m=0}^{\infty} \left|(a_m - b_m)\lambda^n e^{im\pi(j\Delta x)}\right|, \\ &= \sum_{m=0}^{\infty} \left|e^{im\pi(j\Delta x)}\right| \left|a_m - b_m\right| \left|\lambda^n\right|, \\ &= \sum_{m=0}^{\infty} \left|a_m - b_m\right| \left|\lambda\right|^n, \\ &\leq \sum_{m=0}^{\infty} \left|a_m - b_m\right| \sum_{m=0}^{\infty} \left|\lambda\right|^n = \left|V_j^0 - W_j^0\right| \sum_{m=0}^{\infty} \left|\lambda\right|^n.\end{aligned}$$

Hence, if there is to be a K such that $\|\mathbf{V}^n - \mathbf{W}^n\| \leq K\|\mathbf{V}^0 - \mathbf{W}^0\|$, $n\Delta t \leq t_F$, it must be an upper bound on the sum of the $|\lambda|^n$, and so it must be larger than each term in the sum. Suppose that such a K exists for all m. Then $|\lambda|^n \leq K$ and so $n \ln(|\lambda|) \leq \ln(K)$. Suppose that $|\lambda|>1$. Then there will always be some n such that $n > \ln(K)/\ln(|\lambda|)$, and so there will always be some time $t^* = n\Delta t$ at which the stability condition is not fulfilled. The existence of such a t^* may not matter if $n\Delta t > t_F$, but if the model is to be valid for all times then it is a problem.

However, if we can guarantee that $|\lambda| \leq 1$ then the scheme will be stable for all times. From (33), the worst case is for $m = 2$, which gives $|\lambda| \leq 1$ if $-1 \leq 1 - 4\alpha\Delta t/\Delta x^2$, which requires

$$\Delta t \leq \frac{\Delta x^2}{2\alpha}, \tag{36}$$

which is the stability condition on Δt illustrated in section 3.3 for $\alpha = 1$ (figure 1).

This analysis ignores the effects of boundary conditions on stability. Generally any strong non-linearities will require a reduced time step. For example, if radiative boundary conditions were used in the problem described above, the fourth-order polynomial required at the boundaries would produce large errors in the solution if $\Delta t = \Delta x^2/2\alpha$ were used as the time step. The analysis can be extended to more dimensions by considering Fourier series expansions in several directions, and will generally produce an estimate of the form $\Delta t \leq f(\Delta x, \Delta y, \Delta z)$. Further information is available [3, chapters 2, 3 and 5].

For a system of ordinary differential equations of the form, $\mathbf{y}' = \mathbf{f}(\mathbf{y}, t)$ the nature of the solution is partly determined by the Jacobian **J** of the system where $J_{ij} = \partial f_i/\partial y_j$. In particular, the nature of the solution is determined by the nature of the eigenvalues of **J**. For a linear system, these eigenvalues are the independent solutions to the problem, so the solution can be written in the form

$$\mathbf{y} = \sum_{i=1}^{N} \mathbf{a}_i e^{\lambda_i t}.$$

If some of these eigenvalues have negative real parts of very different sizes, so that $\mathrm{Re}(\lambda_i) / \mathrm{Re}(\lambda_j) << 1$, for some $1 \leq i, j \leq N$ with $\mathrm{Re}(\lambda_i)<0$, $\mathrm{Re}(\lambda_j)<0$, then the system is called a **stiff system**. This is a system that involves phenomena occurring over at least two very different time scales. A more general definition applicable to nonlinear systems is “if a numerical method is forced to use, in a certain interval of integration, a step length which is excessively small in relation to the smoothness of the exact solution in that interval, then the problem is said to be stiff in that interval” [10].

Stiff systems will lead to a severe restriction on the time step for stability because, even after the events taking place over the shorter time scale have finished, the stability of the time step will still partly be determined by the short time scale. Often if a physical system consists of many initial high-frequency effects that decay followed by exponential decay over a longer time scale, the system is stiff and so the time step will be severely restricted. If model results appear to oscillate in a region where smooth results are expected despite the use of a small time step, stiffness may be the cause. Generally in such cases it is worth considering an alternative solution method (such as backward differentiation formulae for ordinary differential equations), or a different formulation of the problem so that the short-term effects can be ignored over a longer time scale.

The drawbacks of the test for the stability of a time step are that it becomes more difficult to apply as the model geometry becomes more complicated, and that it can be very difficult to take non-linear effects into account. The stability test is most useful for people writing their own software, particularly for finite difference methods, but it can be important to be aware of the technique and the need for a stable time step when using proprietary software as well. In some cases it is enough to be able to recognise the oscillatory behaviour as shown in figure 1 as characterizing time step instability and correct the model by trial and error, rather than try to identify a suitable stability criterion via analysis.

7.1.3 Mesh convergence tests

Mesh convergence tests check that a mesh is suitable for modelling a specific situation and are based on the assumption that the model will converge to the exact solution. As has been stated above, this convergence depends on the stability and consistency of the method for the situation being considered. Generally the method is to run the same problem with two meshes of different densities and then compare the two sets of results. The simplest form of this method is "halve it and re-run", where a mesh is checked by doubling the mesh density in each direction (halving all element dimensions) and re-running the model to check the results. Similar tests for checking the convergence of a dynamic model can be run by halving the time step and repeating the calculation. Both of these methods are generally very expensive computationally because of the increase in either matrix size or number of time steps.

As an example, consider a flat square plate constrained in all directions around its edges. The plate was meshed with n square 4-noded elements along each edge, and the first 10 natural frequencies of the plate were calculated. Initially, n was set at 2, and for each subsequent mesh n was doubled up to a maximum of $n = 128$. The runs were repeated for n up to 64 using 8-node elements. Some of the calculated frequencies are shown in table 3. Table 4 shows the percentage change in calculated value between each pair of meshes. Figure 30 plots all of the results against degrees of freedom in the model. The degrees of freedom are the unknown values to be calculated during the solution of a model, and the number of degrees of freedom is an indication of the size of a job. The number of degrees of freedom depends on the number of elements and the order of accuracy of the element. Plotting result accuracy against number of degrees of freedom is the best way to combine the results from the two types of element in one plot since it gives an indication of how many values are being used to describe the deformation of the plate.

Several conclusions can be drawn from these tables. The first is that the lower frequencies converge to a value more rapidly than the higher ones. This is to be expected because the higher frequencies require more complex displacements of the structure, and a low-density mesh will not be sufficiently detailed to describe the displacement accurately. The second is that the four-node elements converge more slowly than the 8-node ones. This is expected because the 8-node elements are second-order accurate and so have a higher-order truncation error.

Number of elements per side	Natural frequencies (Hz)							
	First		Fourth		Seventh		Tenth	
	4-node	8-node	4-node	8-node	4-node	8-node	4-node	8-node
2	2.191	2.394	6.799	12.09	13.05	19.45	35.82	25.64
4	2.306	2.381	8.799	9.589	14.47	15.64	23.77	20.45
8	2.357	2.378	9.225	9.550	14.87	15.58	20.08	20.54
16	2.372	2.377	9.461	9.537	15.41	15.57	20.48	20.48
32	2.378	2.377	9.677	9.551	15.95	15.61	21.13	20.60
64	2.378	2.378	9.615	9.574	15.80	15.72	21.19	20.97
128	2.387	-	10.27	-	17.57	-	23.78	-

Table 3: Selected calculated natural frequencies of the plate with different mesh densities.

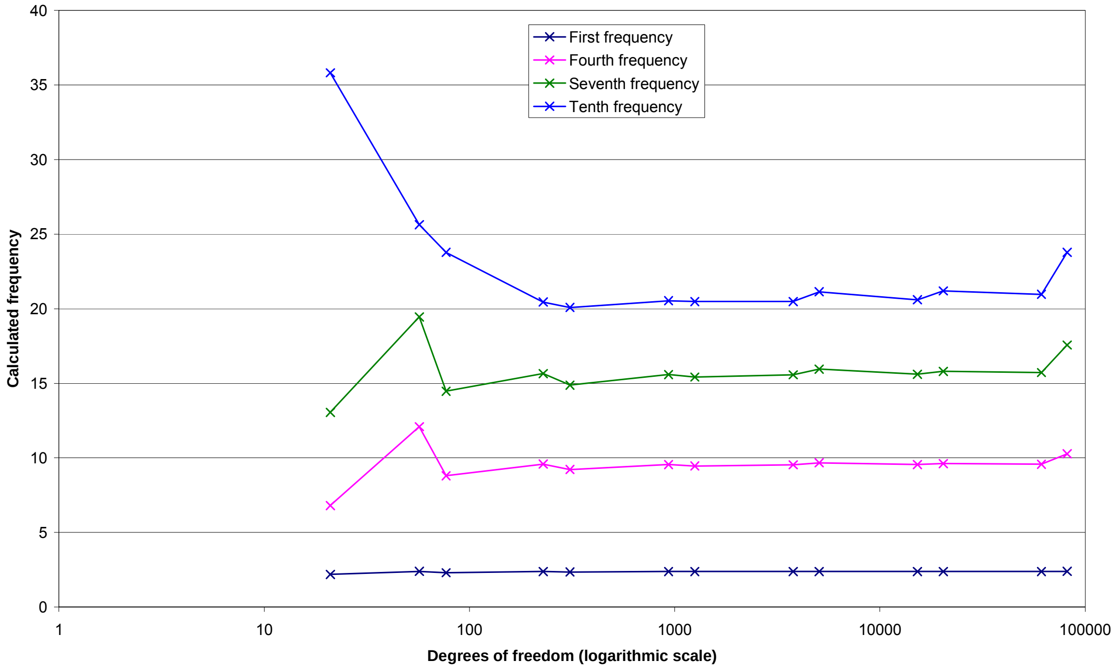

Figure 30: Plot of the results in Table 3 against degrees of freedom in the model used. Note the degrees of freedom are plotted on a logarithmic scale.

Number of elements per side	Percentage change in natural frequencies							
	First		Fourth		Seventh		Tenth	
	4-node	8-node	4-node	8-node	4-node	8-node	4-node	8-node
2→4	4.99	0.55	22.73	26.08	9.81	24.36	50.69	25.38
4→8	2.16	0.13	4.62	0.41	2.69	0.39	18.38	0.44
8→16	0.63	0.04	2.49	0.14	3.5	0.06	1.95	0.29
16→32	0.25	0	2.23	0.15	3.39	0.26	3.08	0.58
32→64	0	0.04	0.64	0.24	0.95	0.7	0.28	1.76
64→128	0.38		6.38		10.07		10.89	

Table 4: Percentage change in calculated natural frequencies when the mesh size is halved.

An interesting point is the seemingly poor performance of the 128 by 128 element mesh. The large percentage changes in the higher frequencies when the number of elements along a side increases from 64 to 128 implies that the solution is heading away from a converged solution. According to the manual for the package used, the accuracy of a natural frequency calculation is related to the ratio of the highest and lowest frequencies, and the use of small elements in the mesh leads to numerical errors in the highest frequency. This poor performance will be caused by the poor condition number of the matrix describing the discretised formulation of the problem. Ideally, only formulations producing well-conditioned matrices should be used. However, generally the full matrix is only formed when the model is solved, and it is not easy to predict the condition of a matrix in advance, so it can be difficult to identify which problems will produce badly conditioned matrices.

In general, the behaviour of a continuous model's results with mesh size can be considered as having three stages, as shown in figure 31. In the first stage, the reduction of the mesh size leads to a reduction in the discretisation error and so the results are converging. In the second stage, the mesh is sufficiently dense that the results have essentially converged to a stable value. In the third, the matrix describing the problem has become ill-conditioned and so the results are becoming unstable and convergence as been lost. The aim of the mesh convergence tests is to try to identify the beginning of the stable region where good results can be obtained quickly. The report "Testing Continuous Modelling Software" [26], addresses the problems of rounding errors more directly and provides a methodology for testing software to see where they become problematic.

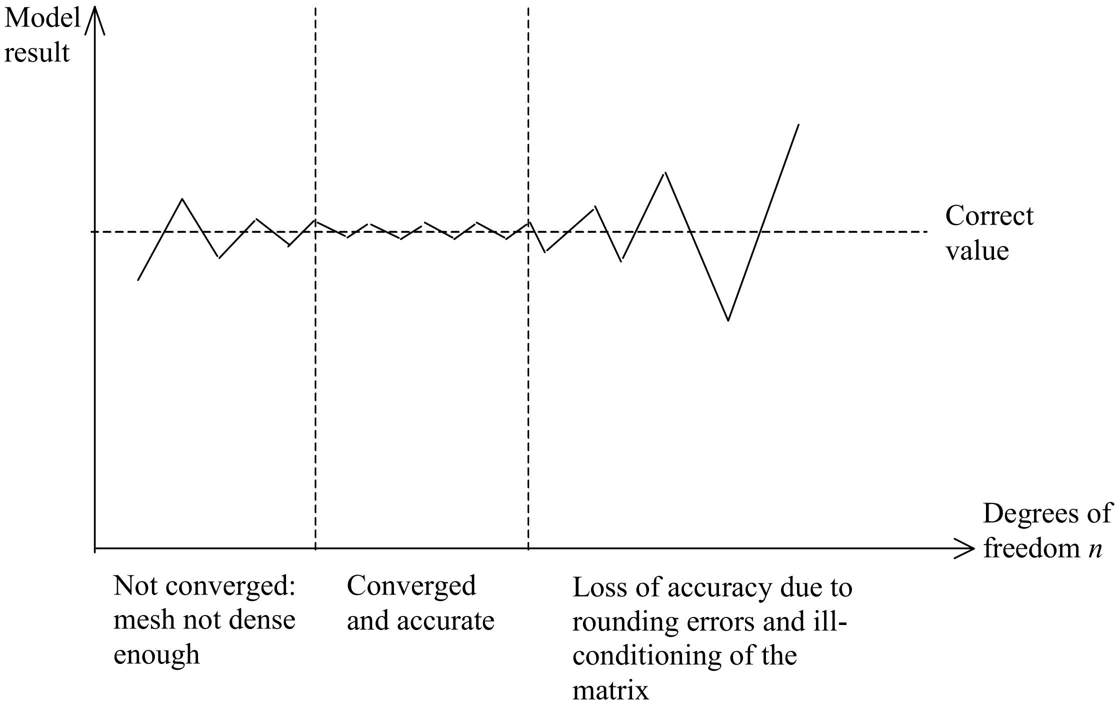

Figure 31: Generalisation of the behaviour of model results with decreasing mesh size.

Mesh refinement

The "halve it and rerun" method can lead to excessive run times, so is often worth increasing the mesh density to a less dramatic degree before rerunning and using a more sophisticated refinement strategy. In general, it is best to refine a mesh in the areas where results are changing most rapidly, for example areas of large deformation or of high temperature gradients. Refining only in areas where the results are of interest is not a good approach, since errors in regions away from the refined mesh could easily affect the results in the refined regions. This spreading of the error is particularly likely in dynamic analyses where errors originating in one place can spread to other areas rapidly. Further advice on mesh improvement and efficient mesh design can be found in the "Guide to the use of finite element and finite difference software" [25].

Mesh refinement may not always lead to a converged solution, for instance if there is some infinite derivative that renders a Taylor series expansion invalid. This lack of convergence is particularly true of models near interfaces and boundaries. It is possible to refine the mesh to a very high level and not obtain a converged solution. For example, if there is a singularity at the boundary it is unlikely that mesh refinement near to the surface will give a good solution. This effect may mean that extra care has to be taken in the mesh design approaching the boundary, and in some cases averaging of results over several elements at a time will give better results than a point-by-point comparison. Finding a solution by averaging is not an ideal method, but since the model is attempting to approximate infinite values, it is understandable that the normal methods of solution fail.

Another potential problem with mesh convergence tests is that if the model inputs (e.g., material properties, boundary conditions, etc.) are changed too much, the mesh may become inadequate. In general, mesh convergence should be re-checked every time a major change is made to the problem definition. With experience, users can spot signs that a mesh is inadequate (the "Guide to the use of finite element and finite difference software" [25] has more detail), but these signs often only occur in the most drastic cases, and a mesh can look reasonable but still produce bad results.

7.2 A posteriori error estimation methods

A posteriori methods use the calculated discrete solution to estimate the error in the solution, in contrast with a priori methods that estimate the solution in terms of knowledge of the exact solution. Since a posteriori methods provide an error estimate from known quantities they are often used in practical calculations, particularly to adapt the discretisation in order to reduce the error.

A full and detailed explanation of the theory underlying these techniques, their application to finite element analysis, and a comprehensive range of error estimators, is available [9]. A less theoretical treatment of some of the simpler techniques is recommended [10, chapters 12 and 14]. An exhaustive survey of work carried out before 2001 on these techniques and their links to adaptive finite element meshes is available [12]. Only a few techniques are given here, but there is a wide range of methods that are suitable for different problems.

A posteriori methods can be split into two main types: the implicit and the explicit. Explicit methods use existing computed solutions and problem data to calculate an error estimator. Implicit methods require the solution of a related problem expressed in terms of the error to obtain an estimate. Implicit methods are generally more time-consuming to use, but often provide tighter bounds on the actual error. Explicit techniques are often faster to use than implicit estimation methods, but they are frequently pessimistic and over-estimate the error because they tend to neglect details that can give error cancellations. Most of the techniques described here are explicit, but a thorough coverage of implicit techniques is available [9].

7.2.1 Spatial methods

In general, continuous modelling problems are discretised by defining a mesh and then assuming a form for the variable of interest (often a localised polynomial) and deriving discrete versions of the differential equations based on these approximate forms. For example, simple linear beam finite elements assume that the displacement on a mesh of nodal points $x_0, x_1, \ldots, x_M$ is of the form

$$\begin{aligned}
u(x) &= \sum_{i=0}^{M} U_i N_i(x), \\
N_i(x) &= \begin{cases} (x - x_{i-1})/(x_i - x_{i-1}), & x_{i-1} \le x \le x_i, \\ (x_{i+1} - x)/(x_{i+1} - x_i), & x_i \le x \le x_{i+1}, \quad i = 1, \ldots, M-1, \\ 0, & \text{otherwise}, \end{cases} \\
N_0(x) &= \begin{cases} (x_1 - x)/(x_1 - x_0), & x_0 \le x \le x_1, \\ 0, & \text{otherwise}, \end{cases} \\
N_M(x) &= \begin{cases} (x - x_{M-1})/(x_M - x_{M-1}), & x_{M-1} \le x \le x_M, \\ 0, & \text{otherwise}, \end{cases}
\end{aligned} \tag{37}$$

(as shown in figure 32), and for more complicated approximations and in higher dimensions, N_i is a higher-order polynomial involving more nodes and more co-ordinates.

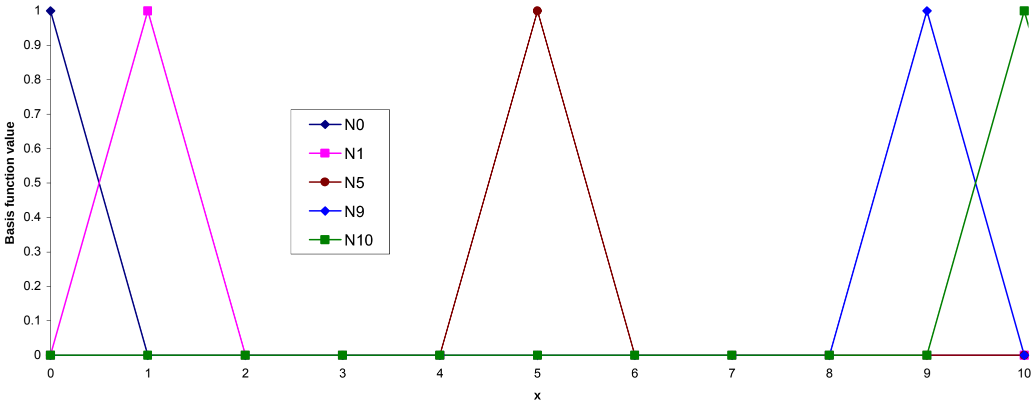

Figure 32: Basis functions N_i for a mesh of one-dimensional linear beam elements using formulae given in (37) with $x_i = i$ and $M = 10$.

Spatial a posteriori methods often use an interpolation or extrapolation based on values calculated using this formulation to provide an improved solution, and then use the difference between the new solution and the original one as an error estimator.

Finite difference methods do not strictly assume a specific form for the solution. Instead they approximate the gradient terms in the continuous formulation of the problem using combinations of discrete point values based on the Taylor expansion (although this formulation will be exact for simple local polynomials, so a form like (37) could apply). This formulation of the discretised problem makes most of the techniques described below unsuitable for application to finite difference problems.

The methods that follow are most commonly applied to finite element problems, although some could be applied to boundary element problems too. The most suitable spatial a posteriori method for finite difference problems is probably Richardson extrapolation from two different meshes, as described in section 7.2.3. Alternatively, if the form of the leading term in the truncation error is known, then finite difference methods, automatic differentiation, or curve fitting techniques could be applied to the calculated solution to estimate the leading term and thus provide a pointwise error estimator. The calculation of this estimator may prove difficult for solutions with only a small change in solution values between adjacent mesh points. The validity of this error estimator also relies on the leading term in the truncation error definitely being the largest term, which may not be the case if higher-order derivatives are very large.

Suppose throughout the following that the continuous problem is of the form

$$\begin{aligned} A\mathbf{u} &= \mathbf{f} \text{ in } \Omega, \\ \mathbf{u} &= \mathbf{g}(\mathbf{x}) \text{ on } \partial\Omega, \end{aligned} \tag{38}$$

where A is a linear differential operator, and that it is to be solved using a finite element approximation. The notation and terminology used in this section are explained in section 2.2.

A typical implicit method would consist of solving a "dual problem" derived from (38). The most general form of the discretised approximation problem can be written as:

Find $\mathbf{U} \in S_h$ such that $\int_\Omega (A\mathbf{U})^{\mathrm{T}} \mathbf{V} d\Omega = \int_\Omega \mathbf{f}^{\mathrm{T}} \mathbf{V} d\Omega$ for all functions $\mathbf{V} \in S_h$,

where S_h is the function space of test functions, for example the space with basis functions as in (37) above. This formulation is commonly written as $B(A\mathbf{U}, \mathbf{V}) = B(\mathbf{f}, \mathbf{V})$. Consider the error **e**, defined as $\mathbf{e} = \mathbf{u} - \mathbf{U}$. Then the aim of the dual problem is to use the adjoint operator A^* to calculate a function φ such that $A^*\varphi = \mathbf{e}$. If such a φ can be found, then $B(\mathbf{e}, \mathbf{e}) = B(\mathbf{e}, A^*\varphi) = B(A\mathbf{e}, \varphi) = B(f - A\mathbf{U}, \varphi)$, which is an estimate of the norm of the error in terms of known quantities. Since the true error is unknown, other quantities are used as estimates of **e** in the calculation of φ. One example is the residual error of the calculated solution, which is a measure of how accurately it satisfies the original differential equation and boundary conditions.

For many common problems such as stress analysis and the heat equation, the operator is self-adjoint, which simplifies matters. Usually the dual problem is solved over disjoint subsets of the domain, since the solution over the whole domain would be computationally equivalent to calculating a new solution. Full details of useful implicit methods are available [9].

The gradient recovery method

One of the simplest techniques for estimating spatial discretisation errors is the gradient recovery method [17]. It is particularly useful for cases where the derivative of the main solution variable is of interest, so for instance it can be used to produce an estimate of the errors in the calculated stress when displacement has been calculated for an elasticity problem, or it can be used to estimate errors in fluxes from boundaries if energy loss rate is of interest.

The formulation in equation (37) leads to continuity of the displacements but not generally of the gradients, which means that for elastic models stress predictions are discontinuous (usually this is true even for higher-order polynomial approximations, although some types of elements do ensure continuity of gradients as well as displacements). This discontinuity means that if the error in the calculated stresses is of interest, a first step towards estimating it may be to calculate an approximation to the stresses that is continuous at the nodes based on the assumption that the stress is continuous in reality. The method is explained more fully by Rencis et al [17].

Suppose some linear function s_{ij} of the derivatives of **u** in (37) is of interest, so that

$$s_{ij} = D_{ijkl} \frac{\partial u_k}{\partial x_l},$$

where D is a constant tensor and the summation convention is used. If the gradient itself is of interest, $D_{ijkl} = \delta_{ik}\,\delta_{jl}$, and for elasticity problems D could be the elasticity tensor. Write the approximate solution from a mesh with n nodes as

$$\mathbf{U}(\mathbf{x}) = \sum_{m=0}^{n} \mathbf{U}^m N_m(\mathbf{x})$$

for some calculated nodal values $\mathbf{U}^i$, and suppose that the approximation to s_{ij} is S_{ij}, which will be a function of the derivatives of the N_m.

Now consider constructing a different approximation T_{ij} to s_{ij}, where

$$T_{ij} = \sum_{m=0}^{n} T_{ij}^m N_m(\mathbf{x}),$$

and the values $T^m{}_{ij}$ to be determined are the values of T_{ij} at the nodes. This approximation will be continuous at the nodes and the $T^m{}_{ij}$ will be chosen so as to make it a higher order of accuracy than S_{ij}. Hence the aim of the technique is to choose the $T^m{}_{ij}$ to minimise the L_2 norm of the difference $S_{ij} - T_{ij}$. This is equivalent [17] to solving the linear algebra problem

$$\sum_{m=0}^{n} A_{rm} T_{ij}^{m} = b_r, \text{ where } A_{rm} = \int_{\Omega} N_m(\mathbf{x}) N_r(\mathbf{x}) d\Omega, \quad b_r = \int_{\Omega} S_{ij} N_r(\mathbf{x}) d\Omega.$$

Then the l_2 norm of $S_{ij} - T_{ij}$ is a first estimate of the L_2 norm of the error in S_{ij}. The A_{rm} and b_r are relatively straightforward to calculate for simple element types and straightforward meshes. It has been found [17] that this technique is quite effective for linear elements, but less so for quadratic and higher orders.

The Superconvergent Patch Recovery method

A slightly more complex technique is superconvergent patch recovery (SPR) [8]. This technique is only applicable to finite element analyses, and relies on a property of the points used for numerical integration within an element. If the results are sampled at these points (called Gauss-Legendre points), then the error in these sampled values is a power of h more accurate than those sampled elsewhere in the element (where h is some characteristic length of the element), provided that certain conditions on the true solution and the discretisation are fulfilled. A survey of these conditions is available [20], but probably the most important condition is that the technique is not generally valid for non-linear problems. The method is also not suitable for problems involving highly localised refinements of the mesh, as these invalidate the conditions on the discretisation.

Suppose that we have elements of order n, so that the error in the primary variable (e.g., displacement, temperature, etc.) is generally $O(h^{n+1})$. If a small patch of elements is grouped together and sampled at the Gauss-Legendre points, indicated by red squares in figure 33 for $n = 1$, then because the error at these points is $O(h^{n+2})$ the primary variable can be approximated within the shaded patch with an expansion of the form

$$u^*(\mathbf{x}) = \sum_{i=1}^{m} a_i N^i(\mathbf{x}),$$

where the N^i are basis functions consisting of powers of x and y, (assuming the problem is two-dimensional), up to total order n+1, and m is $(n + 3)(n + 2)/2$. The coefficients a_i can be found by fitting them to the Gauss-Legendre values in a least-squares sense, so if the values of u at these points are $u^S(\mathbf{x}_i)$, then the problem is to minimise

$$\sum_{i=1}^{k} \left[u^s(\mathbf{x}_i) - u^*(\mathbf{x}_i)\right]^2,$$

where k is the total number of superconvergent points summed over all the elements in the patch. As with the gradient recovery technique described above, this method leads to a linear algebra problem that is straightforward to solve. The difference between this improved approximation and the original calculated value can be used as an error estimate for validation purposes. Often this technique is used to produce improved estimates of gradient quantities, so that the values of the main variable and its derivative are accurate to the same order.

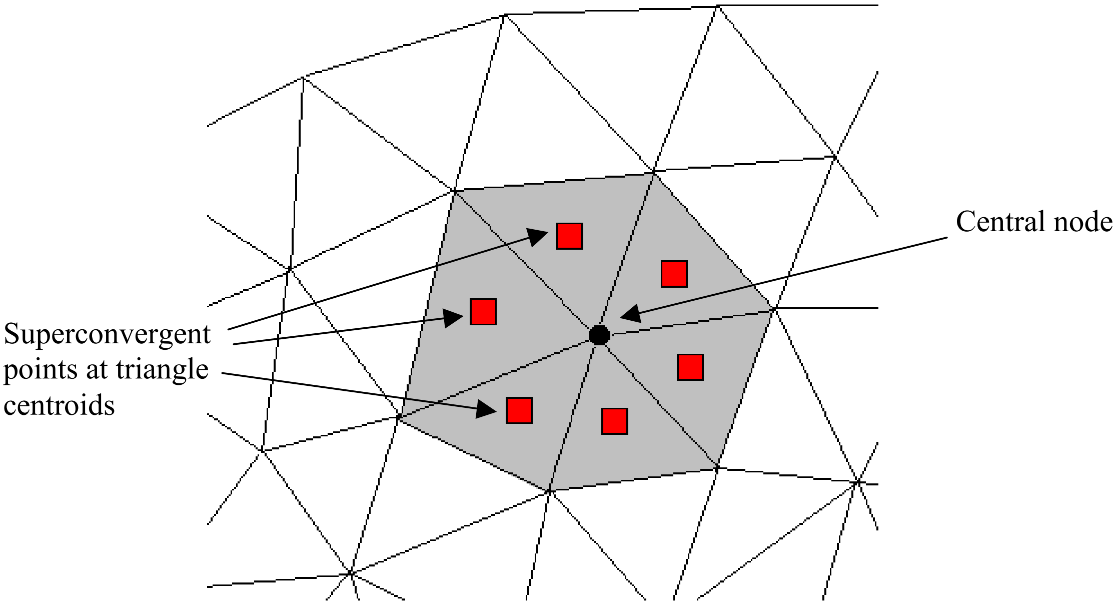

Figure 33: Typical patch used for the superconvergent patch recovery technique. The shaded patch is used to produce an error estimate by interpolating the values calculated at the red squares to provide an improved estimate of the value at the central node.

The most suitable way to choose elements for a patch is to include all elements of which the node of interest is part, as shown in figure 33, since this will use all relevant information that is available. In general, if the number of points used is less than the number of parameters available, the problem will either be degenerate or will not have a solution. This means that sometimes a patch will have to include elements that are not attached to the node in question. For example if a node is on the boundary of a mesh it may not be attached to a sufficient number of elements to form a patch.

As an example, consider the beam shown in figure 34(a). There is an analytic solution that solves an idealised version of this problem, but it is not able to describe the ends of the beam near to the boundary conditions. Instead, a very fine mesh was used to produce a converged solution that was close to the analytic solution away from the ends of the beam, and this was regarded as the true solution to the problem.

This problem was treated as a plane strain problem and was meshed with 1280 triangular elements, as shown in figure 34 (b). The calculated results for σ_{xx} from this mesh were then used in both of the techniques described above to improve their accuracy and estimate their errors. Figure 35 shows the original results and the converged solution that was regarded as accurate.

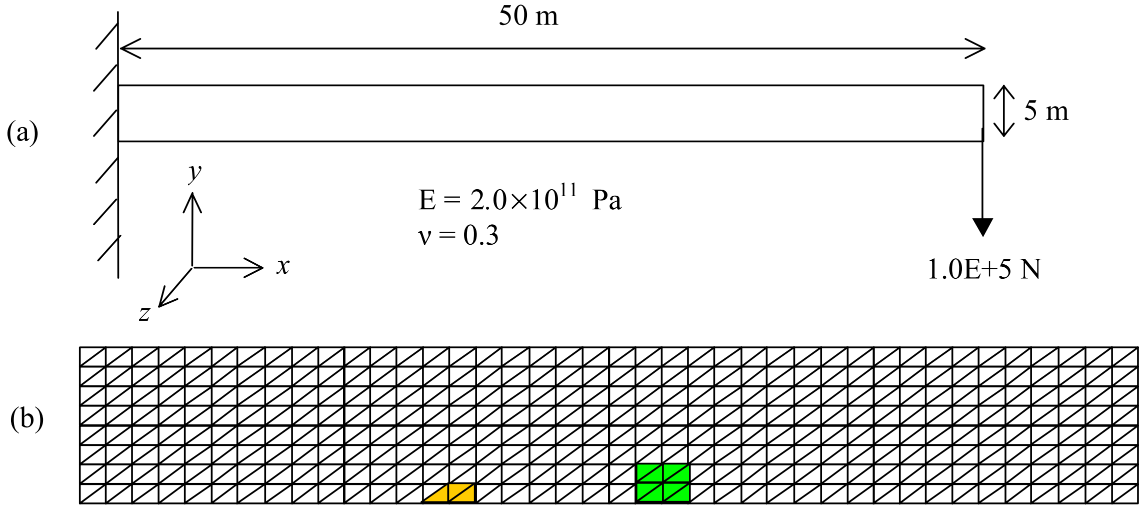

Figure 34: Loaded beam used to demonstrate a posteriori error estimates (a), and mesh used for the initial solution (b). The beam is taken as having unit thickness in the out-of-plane direction. The coloured sections on the beam mesh show a typical three-element patch and eight-element patch as used in the superconvergent patch recovery.

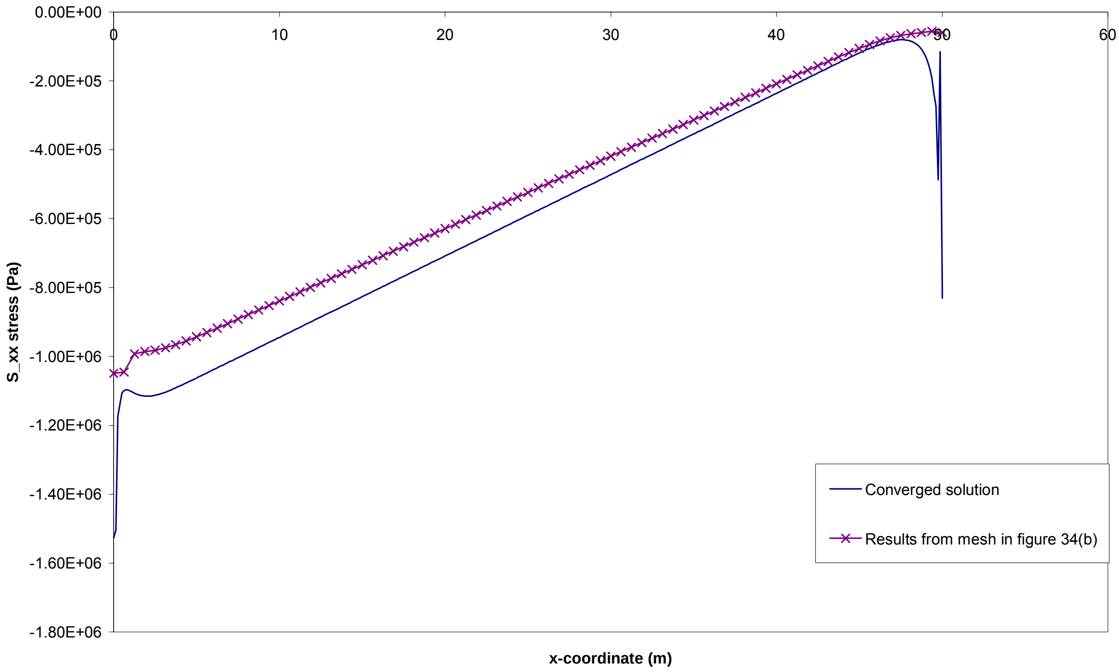

Figure 35: σ_{xx} along the lower surface of the beam, calculated using the converged solution and the mesh in 34(b).

Two different versions of the superconvergent patch recovery were used: one with three elements per patch and one with eight elements per patch. Two example patches are shaded in figure 34(b). The superconvergent patch recovery technique was applied only to the nodes on the lower surface of the beam. Figure 36 shows the pointwise error estimates for σ_{xx} resulting from the calculations.

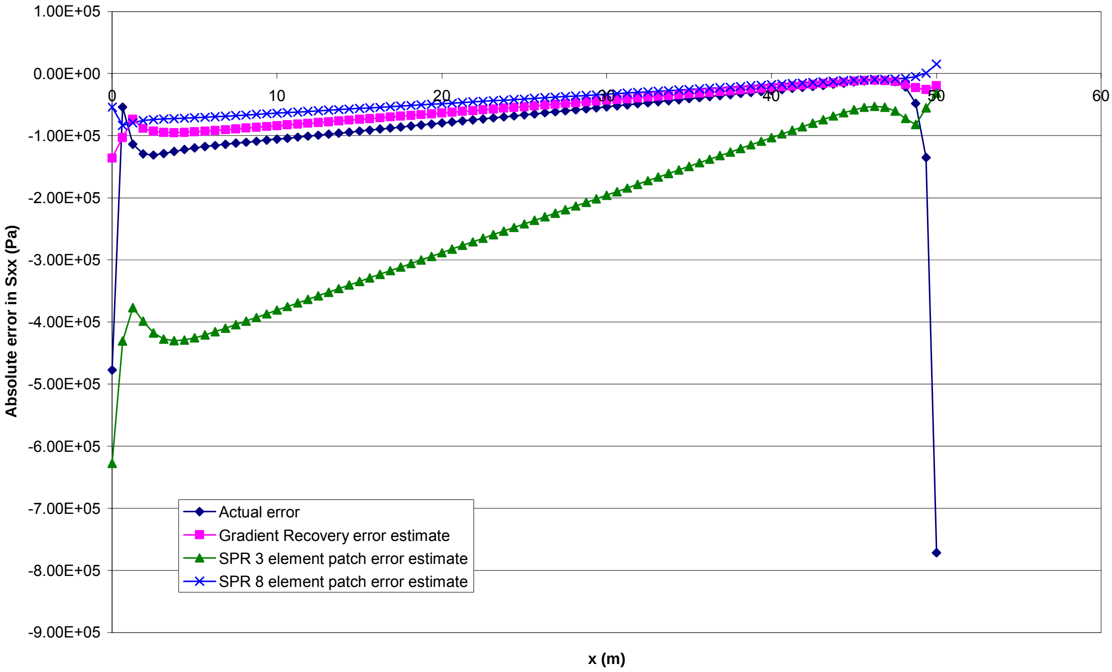

Figure 36: Absolute errors in σ_{xx} along the lower surface of the beam calculated using the mesh in 34(b) (actual error), the gradient recovery technique, and the two versions of the superconvergent patch recovery technique.

It is clear from these results that the choice of patch has a very strong effect on the results and hence the error estimate. Closer examination of the intermediate stages of the calculation showed that this effect is probably due to the three-element patch being unable to describe adequately the variation in the y-direction. This emphasises the importance of choosing an adequate patch.

Figure 36 also shows that both techniques under-estimate the error for most of the length of the beam. A common measure of an error estimate that can be calculated when the true solution is known is the effectivity index

$$\theta_{\text{eff}} = \frac{\left\| e_{\text{estimated}} \right\|}{\left\| e_{\text{actual}} \right\|}.$$

Using the l_2 norm, θ_{eff} was 0.75 for the gradient recovery method, and 0.83 for the superconvergent patch recovery technique with eight elements per patch.

In general, the gradient recovery technique requires the solution of a single large matrix problem, with the matrix size dependent on the number of nodes, whereas SPR requires the solution of a series of smaller matrix problems, one for each node with the matrix size dependent on the order of approximation in the elements and the terms in the matrix dependent on the number of elements in a patch. This difference means that the computational efficiencies will vary according to how many points are of interest, how many nodes there are in total, and how many elements are used per patch. If only a few nodes in a large mesh are of interest, then SPR will be more efficient.

The main problems with the recovery methods described above are that it is difficult to be sure how good the error estimator is, and that for more complex meshes than have been used in the example, the method may be difficult to implement. However, the calculated solution will not become less accurate if these methods are used, and often the output file from finite element packages are of a sufficiently rigid format that a generic program could be written to carry out the method of choice on virtually any problem.

7.2.2 Time step methods

A posteriori error estimation in time is often used to generate adaptive time steps, so that the estimated accuracy of the solution at the previous step (or steps) is used to generate a new time step that is as large as possible, but will still give an accurate stable solution. It is not usual to use such methods for an error estimate at the end of a model run, since error control in time-dependent problems is best done during calculation, as any problems are dealt with more efficiently in terms of run time that way. These methods are included partly for completeness, and partly because they provide a way of including validation and error estimates within the model implementation, even though they are not generally useful for the validation of completed results. An alternative to the methods described in this section is to use a finite element formulation over the time domain and apply the techniques described in the previous section to the problem.

Throughout the following, the methods will be illustrated by application to a single ordinary differential equation of the form $y' = f(y, t)$, but they can be applied to systems of equations and partial differential equations as well.

The predictor-corrector method is probably the most common example of a time step a posteriori method. It involves using two different time stepping difference methods, one explicit and one normally implicit. Explicit methods are those of the form

$$y_{n+1} = F(y_n, y_{n-1}, y_{n-2}, \ldots, y_{n-k}),$$

whereas implicit ones are of the form

$$y_{n+1} = G(y_{n+1}, y_n, y_{n-1}, y_{n-2}, \ldots, y_{n-k}). \tag{39}$$

Explicit methods are quicker to calculate but generally have more limited stability properties, whereas implicit methods are more stable but usually require function inversion due to the presence of y_{n+1} on both sides of the expression (39), and so are usually more time-consuming to calculate. The predictor-corrector technique aims to use a combination of the two methods to provide improved convergence properties and a straightforward calculation.

The predictor-corrector method is a two-step process. The first step is to calculate an initial estimate of the solution at time $t + \Delta t$ using the explicit method (the predictor, $y^P{}_{n+1}$). The second step is to correct this estimate using the implicit method with the estimate $y^P{}_{n+1}$ in G instead of y_{n+1}, which means that the method is no longer being used in an implicit way. The difference between the values from the predictor and the corrector is an error estimate for the predictor value.

Using the notation above, the first step is

$$y^P_{n+1} = F(y_n, y_{n-1}, y_{n-2}, \ldots, y_{n-k}),$$

and the second is

$$y_{n+1} = G\left(y^P_{n+1}, y_n, y_{n-1}, y_{n-2}, \ldots, y_{n-k}\right).$$

The method can be extended by the repeated use of calculated estimates in (39), so that

$$
\begin{aligned}
y_{n+1}^{A} &= F(y_n, y_{n-1}, y_{n-2}, \ldots, y_{n-k}),\\
y_{n+1}^{B} &= G\left(y_{n+1}^{A}, y_n, y_{n-1}, y_{n-2}, \ldots, y_{n-k}\right),\\
y_{n+1}^{C} &= G\left(y_{n+1}^{B}, y_n, y_{n-1}, y_{n-2}, \ldots, y_{n-k}\right),\\
&\vdots\\
y_{n+1} &= G\left(y_{n+1}^{M}, y_n, y_{n-1}, y_{n-2}, \ldots, y_{n-k}\right).
\end{aligned}
$$

This is called a P(EC)^{M} method (P for prediction, E for evaluation, C for correction) and is effectively using functional iteration to find a solution for y_{n+1} in (39) without having to invert G explicitly. The truncation error of this method will be of the form

$$C_{n+1}h^{n+1} + D_{r+1}h^{r+M+1} + O\left(h^{n+2}\right) + O\left(h^{r+M+2}\right),$$

where the truncation error of the predictor method is $D_{r+1}h^{r+1} + O(h^{r+2})$ and that of the corrector is $C_{n+1}h^{n+1} + O(h^{n+2})$. Most useful methods have $M \geq 1$ so the truncation error is an improvement on that of the explicit method, provided that the implicit method is of higher order than the explicit one.

As an example, consider the simple case where the predictor is the forward Euler method,

$$y_{n+1}^{P} = y_n + hf(y_n, t_n),$$

and the corrector is the (implicit) trapezoidal rule

$$y_{n+1} = y_n + \frac{h}{2}\left[f(y_n, t_n) + f\left(y_{n+1}^{P}, t_{n+1}\right)\right].$$

The predictor has a truncation error with a leading term that has $r = 1$, and the corrector has $n = 2$. This means that the overall truncation error should be $O(h^2)$ as well, and this has been shown by trials with different step sizes.

Figure 37 shows the logarithms of the errors obtained when applying the method to the problem

$$y' = -\frac{y^4}{3}, \quad y(0) = 1, \qquad (40)$$

which has analytic solution $y(t) = (1 + t)^{-1/3}$ The figure indicates that the predictor-corrector method has similar errors to the trapezoidal method, but the implicit trapezoidal method used alone takes several iterations of a Newton-Raphson solver to obtain the results, which makes it less computationally efficient than the predictor-corrector where all calculations are explicit.

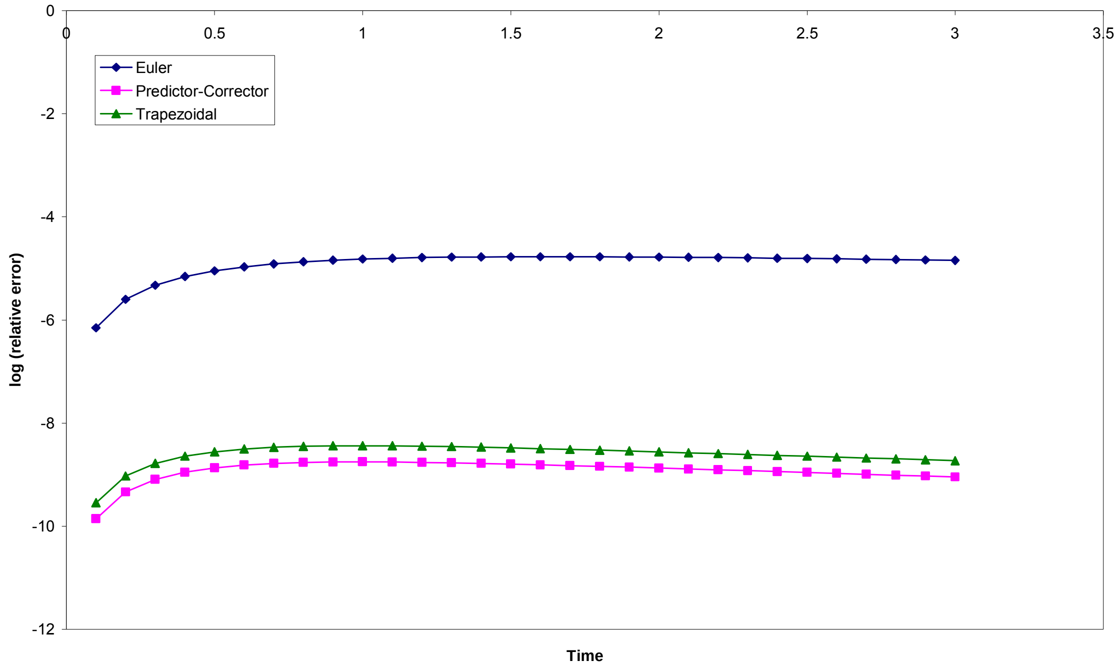

Figure 37: Plot of the logarithm of local errors versus time obtained when using forward Euler, trapezoidal, and predictor-corrector methods on the problem (40).

Another example of a posteriori time step methods is the use of Richardson extrapolation (see section 7.2.3) in time to provide error estimates leading to adaptive time stepping. For instance, if the solution at t_k is known, and the solutions at $t_k + \Delta t$ and $t_k + \Delta t / 2$ are calculated, then error estimates can be obtained for these two solutions using the method given in 7.2.3. The error estimates can then be compared with some tolerance, and the time step can be chosen according to the flow chart in figure 38. All the drawbacks listed in section 7.2.3 will apply to this method, in particular the requirement of knowing the order of the truncation error.

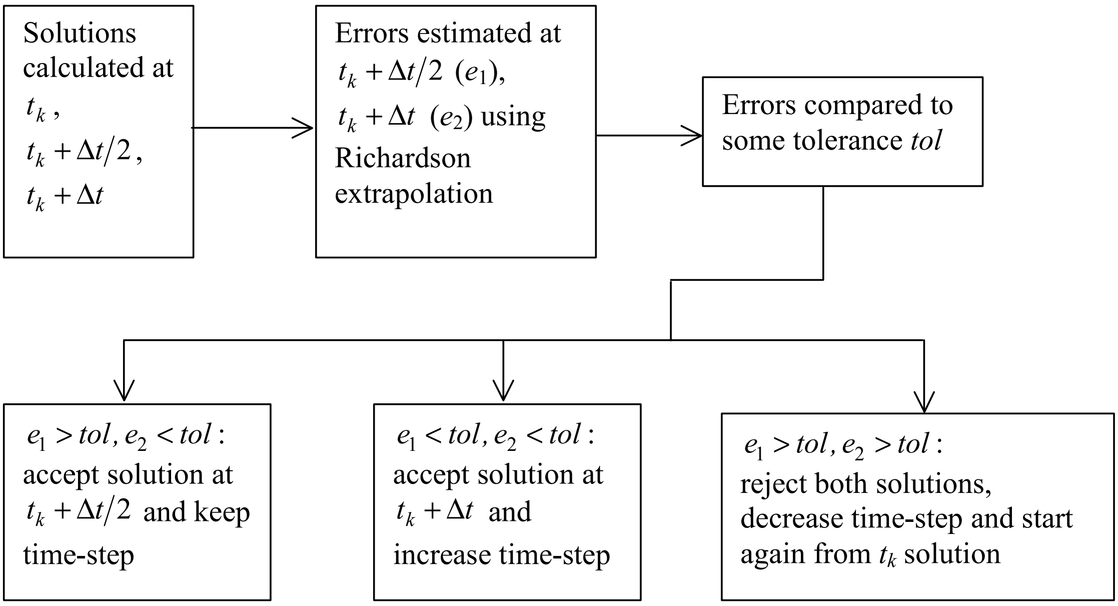
Solutions calculated at t_k, $t_k + \Delta t/2$, $t_k + \Delta t$
Errors estimated at $t_k + \Delta t/2$ (e_1), $t_k + \Delta t$ (e_2) using Richardson extrapolation
Errors compared to some tolerance *tol*
$e_1 > tol, e_2 < tol$: accept solution at $t_k + \Delta t/2$ and keep time-step
$e_1 < tol, e_2 < tol$: accept solution at $t_k + \Delta t$ and increase time-step
$e_1 > tol, e_2 > tol$: reject both solutions, decrease time-step and start again from t_k solution

Figure 38: Adaptive time stepping algorithm based on Richardson extrapolation for the error estimate.

7.2.3 Richardson extrapolation

Richardson extrapolation is a way of using two sets of calculated results to produce either an error estimate or an improved estimate of the actual result. It relies on having two sets of results for the same problem with differing mesh sizes, and can be used for error estimation in space or time.

The procedure as outlined here is very general, and is based on the same ideas as Romberg integration, a numerical integration technique [24, section 4.3]. It is commonly used as the error estimator in adaptive quadrature techniques. These techniques estimate the local error of a numerical approximation to an integral in each of the subdivisions on which the approximation was calculated, and then calculate a global error estimate in terms of these local estimations. Most techniques then alter the subdivision to reduce the error to within some given tolerance. The predictor-corrector technique described in the previous section is very similar to these techniques.

Suppose that the result of interest I^* is approximated by a function of the mesh size $I(h)$, where $I^* = I(0)$ so that $I(h)$ is a consistent approximation to I^*. If it is known that the approximation method is such that

$$I(h) = I^* + h^k \left.\frac{\partial^n I}{\partial x^n}\right|_{x=0} + O\left(h^{k+1}\right), \qquad (41)$$

then

$$I\left(\frac{h}{r}\right) = I^* + \frac{h^k}{r^k}\left.\frac{\partial^n I}{\partial x^n}\right|_{x=0} + O\left(\frac{h^{k+1}}{r^{k+1}}\right)$$

for some $r > 1$ so that the second mesh is a refinement of the first. Hence

$$I(h) - r^k I\left(\frac{h}{r}\right) = \left(1 - r^k\right) I^* + O\left(h^{k+1}\right),$$

and so a better estimate of I^* is

$$\frac{I(h) - r^k I\left(\dfrac{h}{r}\right)}{1 - r^k}, \qquad (42)$$

and error estimates for the two sets of results can be calculated from this estimate. As well as being applied to all the results of a model as is shown in the example below, this method could be used to improve local estimates of a quantity at a specific point following local mesh refinement. This local improvement may need care since it is likely that $I(h)$ will be affected by results away from the point of interest, so the dependence on local mesh refinement may not lead to $I^* = I(0)$ and convergence fails.

As an example, consider the results plotted in figure 29. Equation (26) states that the truncation error is proportional to Δx, so $k = 1$ in (41). Using the results for $\Delta x = 0.01$ and $\Delta x = 0.005$ in (42) gives the results shown in figure 39.

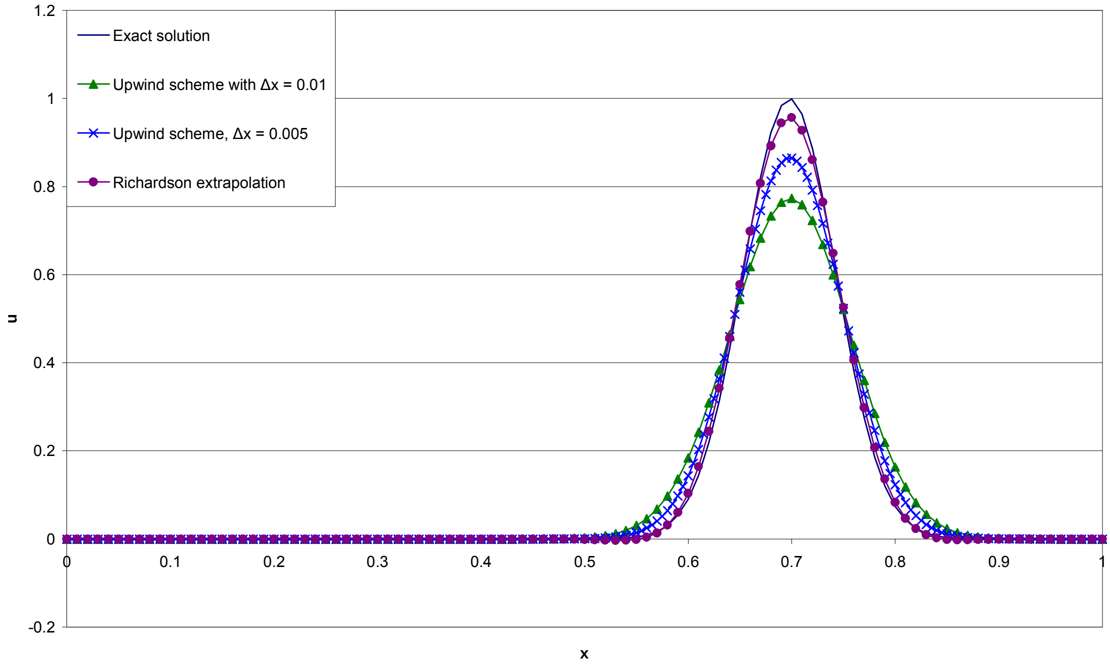

Figure 39: Analytic solution to (23), calculated results for $\Delta x = 0.01$ and $\Delta x = 0.005$, and their Richardson extrapolation.

The main drawback of this technique is the requirement for a Taylor expansion to exist. If the problem involves discontinuities such as shocks or singularities, the Taylor expansion will be invalid in the region of the shock, so the new approximation will not be valid there. Additionally, amplification of round-off errors can occur. Consider an additional term in equation (41), so that

$$I(h) = I^* + h^k \left.\frac{\partial^n I}{\partial x^n}\right|_{x=0} + h^{k+1} \left.\frac{\partial^{n+1} I}{\partial x^{n+1}}\right|_{x=0} + O\left(h^{k+2}\right),$$

$$I\left(\frac{h}{r}\right) = I^* + \frac{h^k}{r^k} \left.\frac{\partial^n I}{\partial x^n}\right|_{x=0} + \frac{h^{k+1}}{r^{k+1}} \left.\frac{\partial^{n+1} I}{\partial x^{n+1}}\right|_{x=0} + O\left(\frac{h^{k+2}}{r^{k+2}}\right),$$

and hence

$$I(h) - r^k I\left(\frac{h}{r}\right) = \left(1 - r^k\right) I^* + h^{k+1}\left(1 - \frac{1}{r}\right) \left.\frac{\partial^{n+1} I}{\partial x^{n+1}}\right|_{x=0} + O\left(h^{k+2}\right),$$

so that if

$$\frac{1}{r^k} \left.\frac{\partial^n I}{\partial x^n}\right|_{x=0} < h\left(1 - \frac{1}{r}\right) \left.\frac{\partial^{n+1} I}{\partial x^{n+1}}\right|_{x=0}, \qquad (43)$$

then the leading term in the truncation error has increased, so the new approximation may be worse. For example, using the trapezium method to evaluate the integral of $\sin\theta$ between 0 and π gives a value of 1.9835 using $\Delta\theta = \pi/10$ and 1.9959 using $\Delta\theta = \pi/20$, so the errors are 0.0165 and 0.0041 respectively. If these values are used to calculate a Richardson extrapolation, the value obtained is 2.0082 which is worse than the previous best value, because the condition (43) is fulfilled.

Another problem is that if the exact form of the truncation error is not known, k is not known and so identifying the correct value to be used in (42) may not be possible. This need for knowledge of k makes usage of the technique on finite element solutions problematic, since sometimes neither h nor k are well-defined for a non-uniform mesh.

Some references, such as Umar et al [11], use a more generalised reliance of the error on the number of elements used in the model to extrapolate to improved results. Care must be taken when doing this, since the number of elements present in a mesh does not characterise the mesh behaviour very well. It would be easy to produce two meshes with the same number of elements, but radically different error bounds. All the examples given by Umar et al [11] involve global uniform refinements of the original mesh, which contributes to the success of the method outlined there. In general, if Richardson extrapolation is to be used for global error estimation, the mesh and refinements should as uniform as possible.

8 Multiple validation methods: a worked example

This section gives a worked example of the application of multiple validation methods to a single problem. This demonstrates how using the different tests together can validate many different aspects of a model.

8.1 The problem

The problem is a model of an experiment that measures thermal diffusivity using the laser flash method. A uniform cylinder of material is prepared and is held in place in a vacuum by three pins. One circular face of the cylinder is heated with a laser flash, and the resulting temperature rise of the opposite face of the cylinder is measured. This process is illustrated in figure 40.

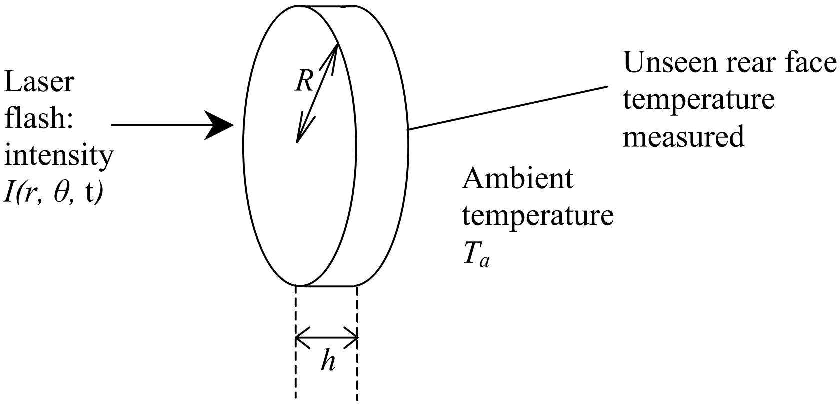

Figure 40: Illustration of the laser flash experiment for measuring thermal diffusivity.

A model describing the heat transfer processes within this experiment is required. It will include the heat input due to the laser, radiative heat losses from the faces of the sample, and conduction within the sample. It is assumed that the convective and conductive losses from the sample are negligible, as it is in vacuo and is supported by pins.

The appropriate continuous model is the heat equation in cylindrical polar coordinates, with non-linear boundary conditions to model the radiative heat losses. The full mathematical formulation is

$$\frac{\partial}{\partial t}\left(\rho c_p T\right) = \nabla.(\lambda \nabla T) + \delta(z) I(r,\theta,t), \quad 0 \le r \le R,\ 0 \le z \le h,\ 0 \le \theta \le 2\pi,\ 0 \le t \le t_0, \quad (44)$$

$$T(r,\theta,z,0) = T_a,$$

$$\lambda \left.\frac{\partial T}{\partial r}\right|_{r=R} = -\sigma \varepsilon_{\text{hem}} \left(T^4(R,\theta,z,t) - T_a^4\right),$$

$$\lambda \left.\frac{\partial T}{\partial z}\right|_{z=0} = \sigma \varepsilon_{\text{hem}} \left(T^4(r,\theta,0,t) - T_a^4\right),$$

$$\lambda \left.\frac{\partial T}{\partial z}\right|_{z=h} = -\sigma \varepsilon_{\text{hem}} \left(T^4(r,\theta,h,t) - T_a^4\right),$$

where T is the temperature at a point, (r, θ, z) are the cylindrical polar coordinates, T_a, R and h are as shown in figure 40, I is the laser intensity in Wm^{-2}, and λ, ρ, ε_{hem}, c_p, and σ are the various thermal properties and constants that define the material's behaviour.

Since the sample is cylindrical, cylindrical polar coordinates are used in the mesh as they can describe the geometry more easily than Cartesian coordinates. The chosen discretisation method is finite differences, for ease of programming. If the equation (44) is written out in full in cylindrical polar coordinates, it gives

$$\frac{\partial}{\partial t}\left(\rho\, c_p T\right)=\frac{\partial}{\partial r}\left(\lambda\frac{\partial T}{\partial r}\right)+\frac{1}{r}\lambda\frac{\partial T}{\partial r}+\frac{1}{r^2}\frac{\partial}{\partial \theta}\left(\lambda\frac{\partial T}{\partial \theta}\right)+\frac{\partial}{\partial z}\left(\lambda\frac{\partial T}{\partial z}\right)+\delta(z)I(r,\theta,t),$$

which is problematic as $r \to 0$. To avoid this problem, around $r = 0$ the differential equation is written as

$$\frac{1}{\alpha}\left.\frac{\partial u}{\partial t}\right|_{r=0}=\left.\frac{\partial^2 u}{\partial z^2}\right|_{r=0}+\frac{4}{\Delta r}\int_0^{2\pi}\left.\frac{\partial u}{\partial r}\right|_{r=\Delta r/2}d\theta$$

obtained by considering a small cylinder of radius Δr / 2 centred on $r = 0$. This has led to two types of discrete equation being developed: one of the form

$$\begin{aligned}\frac{U_{ijk}^{n+1}-U_{ijk}^{n}}{\alpha\Delta t}&=\frac{U_{i+1jk}^{n}-2U_{ijk}^{n}+U_{i-1jk}^{n}}{\Delta r^2}+\frac{1}{r_i}\frac{U_{i+1jk}^{n}-U_{i-1jk}^{n}}{2\Delta r}\\&+\frac{1}{r_i^2}\frac{U_{ij+1k}^{n}-2U_{ijk}^{n}+U_{ij-1k}^{n}}{\Delta\theta^2}+\frac{U_{ijk+1}^{n}-2U_{ijk}^{n}+U_{ijk-1}^{n}}{\Delta z^2}+\frac{I_{ijk}^{n}}{\lambda}+f\left(U_{ijk}^{n},T_a\right),\end{aligned}\qquad(45)$$

for $r \neq 0$, where f represents radiative heat loss if required, and one of the form

$$\begin{aligned}\frac{U_{00k}^{n+1}-U_{00k}^{n}}{\alpha\Delta t}&=\left\{\frac{4}{\Delta r^2}\left[\frac{1}{J+1}\sum_{j=0}^{J}U_{1jk}^{n}-U_{00k}^{n}\right]+\left[\frac{U_{00k+1}^{n}-2U_{00k}^{n}+U_{00k-1}^{n}}{\Delta z^2}\right]\right\}\\&+\frac{I_{00k}^{n}}{\lambda}+f\left(U_{00k}^{n},T_a\right)\end{aligned}\qquad(46)$$

around $r = 0$.

8.2 Validation of the model

Seven methods have been identified as being suitable for validation of this model. They are:

- Visual inspection of results
- Conservation of energy
- Consistency test
- Stability test
- Comparison with analytic solution
- Comparison with an equivalent model
- Comparison with experimental data

Several of these methods (particularly the a priori ones) were implemented during the development process but they will be reiterated here to show their use.

A posteriori methods were not used in this case since they either require multiple runs with different meshes or adaptive time stepping, and it was decided that these options were too time-consuming and complex for the initial development.

The choice of methods above should validate most aspects of the model. The main feature that will not be validated is the non-uniform laser profile definition. This is due to a lack of suitable data. When suitable experimental results are available, the comparison with experiment test will be re-run taking the non-uniformity into account. If more detail was required about the importance of the laser profile, a sensitivity analysis could be carried out to establish how the I affects the model results.

Multiple solutions have been considered, but by considering the alternative and reaching a contradiction, it can be shown that (44) has a unique solution for a given set of boundary conditions, and since (45) is explicit and linear it is likely to have a unique solution, so non-uniqueness is not a great concern.

8.2.1 Visual inspection of results

Three main aspects of the model can be checked visually. The first aspect is that as $t \rightarrow \infty$ all the heat should be lost to the atmosphere and so all the temperature values should be T_a. The second is that the overall shape of the rear-face measurement curve is well understood from experimental experience: it should rise sharply, turn, and decay gradually. The third is that if the laser intensity profile is uniform, then the temperature results should be wholly axisymmetric since material uniformity is assumed and the boundary conditions are axisymmetric.

The first test was carried out for a single typical set of input data, and the model gave the expected result a sufficiently long time. The second and third tests are checked after every model run. Typical examples of the results after post-processing are shown in figure 41. No attempt has been made to quantify the extent to which these results "look right", but they do behave in the expected manner.

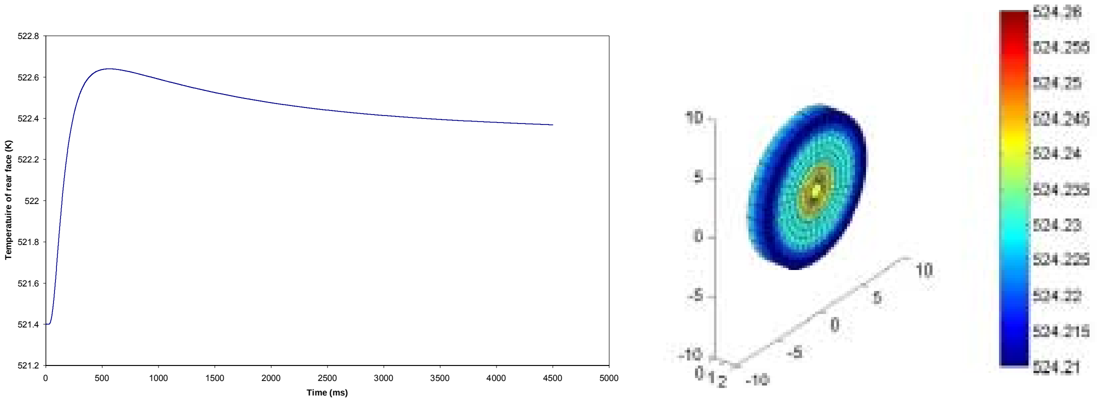

Figure 41: Typical output of the post-processing, showing a temperature curve that rises then dies away and a radially symmetric temperature profile, as expected.

This method has tested the broad properties and behaviour of the model without looking at the numerical accuracy of any features.

8.2.2 Conservation of energy

If the radiative heat loss is ignored by setting $\varepsilon_{\text{hem}} = 0$, then energy will be conserved in the sample. Thus, the expected temperature rise in the sample due to a laser pulse of known energy can be calculated and compared with the model results. The laser energy E will raise the sample temperature by $E / (Mc_p)$ where M is the sample mass and c_p is the specific heat capacity of the material. The laser energy is the product of the laser intensity, the sample surface area, and the laser pulse duration.

The results at 1000s were as follows:

Mass of sample (kg)	4.26×10^{-4}	Total energy in (J)	0.879
Specific heat capacity ($\mathrm{Jkg^{-1}\,K^{-1}}$)	2000	Expected temperature rise (K)	1.033
Specimen surface area ($\mathrm{m^2}$)	1.13×10^{-4}	Model initial temperature (K)	521.4000
Laser pulse intensity ($\mathrm{W/m^2}$)	7.77×10^{-4}	Model final temperature (K)	522.4268
Laser pulse duration (s)	0.1	Model temperature rise (K)	1.027

Table 5: Calculation of expected temperature rise due to known laser energy input

Hence the computer model results and the calculation were in agreement to three significant figures. The disparity between the results is due to the model run time still being insufficient for the temperature to be perfectly uniform within the sample, meaning that the rear face is at a slightly lower temperature than the front face.

This method has checked numerically that the model has an expected overall property of the physical system. This property was also a property of the continuous model, so this is an external consistency check and an inter-model consistency check.

8.2.3 Consistency test

The truncation error of the discretised equations in (45) and (46) can be calculated to ensure that the method and the partial differential equation are consistent with one another. The full algebra will not be given here, but the appropriate equations are

$$\begin{aligned}
&\frac{u^{n+1}_{ijk}-u^{n}_{ijk}}{\alpha\Delta t}-\frac{u^{n}_{i+1jk}-2u^{n}_{ijk}+u^{n}_{i-1jk}}{\Delta r^{2}}-\frac{1}{r_{i}}\frac{u^{n}_{i+1jk}-u^{n}_{i-1jk}}{2\Delta r}\\
&-\frac{1}{r_{i}^{2}}\frac{u^{n}_{ij+1k}-2u^{n}_{ijk}+u^{n}_{ij-1k}}{\Delta\theta^{2}}-\frac{u^{n}_{ijk+1}-2u^{n}_{ijk}+u^{n}_{ijk-1}}{\Delta z^{2}}=\\
&\frac{1}{\alpha}\frac{\Delta t}{2}\frac{\partial^{2}u}{\partial t^{2}}-\frac{\Delta r^{2}}{12}\left(\frac{\partial^{4}u}{\partial r^{4}}+\frac{4}{r}\frac{\partial^{3}u}{\partial r^{3}}\right)-\frac{r^{2}\Delta\theta^{2}}{12}\frac{\partial^{4}u}{\partial\theta^{4}}-\frac{\Delta z^{2}}{12}\frac{\partial^{4}u}{\partial z^{4}}+O\left(\Delta t^{2},\Delta r^{4},\Delta\theta^{4},\Delta z^{4}\right),
\end{aligned}\tag{47}$$

and

$$\begin{aligned}
&\frac{u^{n+1}_{00k}-u^{n}_{00k}}{\alpha\Delta t}-\left\{\frac{4}{\Delta r^{2}}\left[\frac{1}{J+1}\sum_{j=0}^{J}u^{n}_{1jk}-u^{n}_{00k}\right]+\left[\frac{u^{n}_{00k+1}-2u^{n}_{00k}+u^{n}_{00k-1}}{\Delta z^{2}}\right]\right\}=\\
&\frac{\Delta t}{2\alpha}\left.\frac{\partial^{2}u}{\partial t^{2}}\right|_{r=0}-\left\{\frac{1}{J+1}\sum_{j=0}^{J}\frac{\Delta r}{6}\left.\frac{\partial^{3}u}{\partial r^{3}}\right|_{r=\Delta r/2}+\frac{\Delta z^{2}}{12}\left.\frac{\partial^{4}u}{\partial z^{4}}\right|_{r=0}\right\}+O\left(\Delta t^{2},\Delta r^{3},\Delta z^{4}\right).
\end{aligned}\tag{48}$$

These expressions show that the truncation error at the centre, in (48), is of a lower order in the radial mesh size than elsewhere due to the need to be careful around $r = 0$. However, since the remainders on the right hand sides of (47) and (48) both tend to zero as max $[\Delta t, \Delta r, \Delta\theta, \Delta z] \to 0$, the discrete method is consistent with the problem.

This consistency test has checked that the solution of the discrete version of the model will converge to the solution of the continuous model as the step sizes tend to zero.

8.2.4 Stability test

An approximate stability criterion for the method can be calculated using the techniques outlined in section 7.1.2. As the technique used there is linear stability analysis, it is not suitable for cases where the boundary conditions are non-linear, as is the case in this problem due to the radiative heat loss, which is fourth-order in temperature. However, during the experiment the temperature of the sample and the ambient temperature do not generally differ by more than two degrees, so the boundary conditions are sufficiently close to linear that linear stability analysis will be satisfactory.

Applying the technique in section 7.1.2 shows that if the boundary conditions are taken to be linear, the restriction on the time step is

$$-2 \le \alpha\Delta t\left[I_{klm}^{n} - 4\left\{\frac{1}{\Delta r^2}\left(1+\frac{1}{k^2\Delta\theta^2}\right)+\frac{1}{\Delta z^2}\right\}\right] \le 0 \quad \forall\, k,l,m,n,$$

where I^n_{klm} is the contribution from the laser. Since $I^n_{klm} \ge 0$ throughout (since the laser is generating heat rather than removing it) the worst case of the lower bound will be when $I^n_{klm} = 0$ and $k = 1$, which leads to a bound on the time step of

$$\frac{\Delta r^2 \Delta z^2 \Delta\theta^2}{2\left\{\Delta z^2\left(\Delta\theta^2+1\right)+\Delta\theta^2\Delta r^2\right\}} \ge \alpha\Delta t.$$

The upper bound will be exceeded if

$$I_{klm}^{n} > 4\left\{\frac{1}{\Delta r^2}\left(1+\frac{1}{k^2\Delta\theta^2}\right)+\frac{1}{\Delta z^2}\right\}$$

for some *k, l, m, n.* If this is the case, the problem will have to be re-run with a finer spatial mesh since this criterion is independent of the time step. Additionally, this criterion may mean that the temperature rise would be such that radiative losses would become more significant, and so the calculated time step would be unstable anyway and non-linear stability analysis methods would be needed.

The problem under examination is one for which an adaptive time stepping method using a posteriori methods may be useful, particularly since the laser flash only provides heat input at the start of the run time. Such a method has not yet been implemented.

This stability test identifies a stable time step and completes the check that the solution of the discretised problem converges to the solution of the continuous problem as the step size tends to zero.

8.2.5 Comparison with analytic solution

Analytic solutions to this problem exist under certain simplifying assumptions [19]. These assumptions are that the laser flash is instantaneous and spatially uniform, and that there is no radiative heat loss. Under these assumptions, the solution can be written in terms of an infinite sum. Using results from the same model run as was used in section 8.2.2, the temperature at the centre of the rear face of the sample was compared with the analytic solution.

The results and the errors are shown in figures 42 and 43. Summed over 2000 points in time the l_2 error was 0.025 K. Further comparison of the two graphs shows that the error shown in figure 43 is approximately proportional to the second derivative of the analytic solution with respect to time. This is what would be expected from the form of the truncation error as in (48).

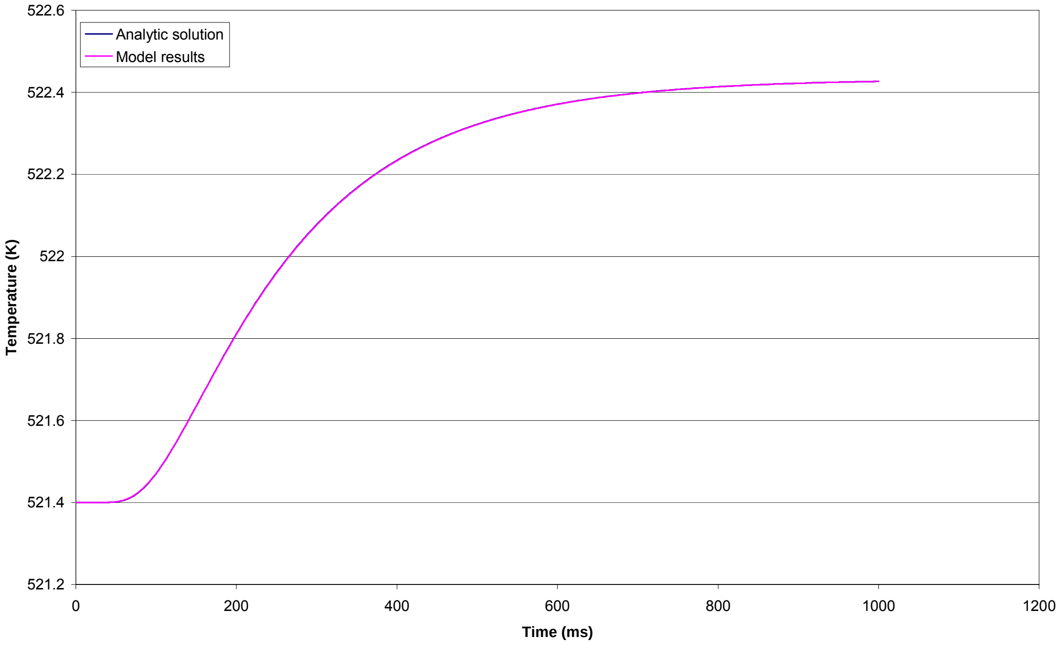

Figure 42: The model results and the analytic solution for a case with an instantaneous uniform laser pulse and no radiative cooling.

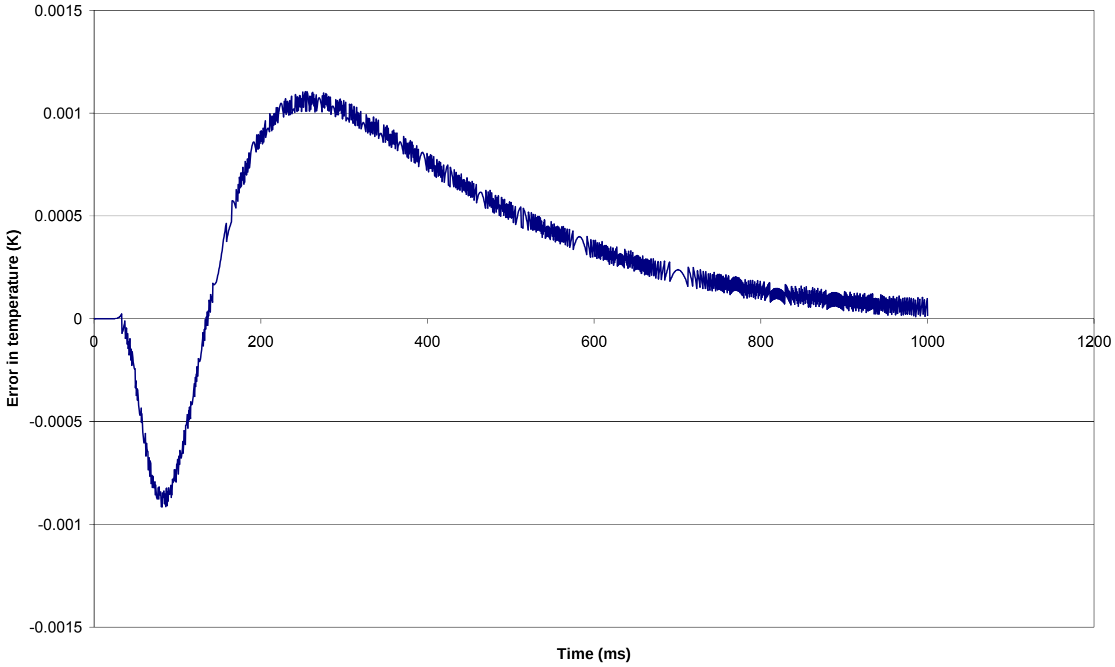

Figure 43: Difference between the analytic solution and model results shown in figure 42.

This comparison with an analytic solution checks the details of the model against a solution to a simplified version of the continuous problem, which tests the software implementation as well as the approximation method.

8.2.6 Comparison with an equivalent model

The equivalent model chosen was an axisymmetric simplification of (44). This has no dependence on θ and so runs considerably faster as it has a fraction of the number of nodes. The resulting equation is

$$\frac{\partial}{\partial t}\left(\rho\, c_p T\right)=\frac{\partial}{\partial r}\left(\lambda\frac{\partial T}{\partial r}\right)+\frac{1}{r}\lambda\frac{\partial T}{\partial r}+\frac{\partial}{\partial z}\left(\lambda\frac{\partial T}{\partial z}\right)+\delta(z)I(r,t),$$

which can be solved as a two-dimensional problem. This problem was solved using the same number of nodes in the r and z directions as had been used in the full three-dimensional model.

Figure 44 shows the difference between the results of the two models. Summed over 5000 points these differences had an l_2 norm of 0.093, which is good agreement. It is not clear where the differences between the models arise, since the material properties, boundary conditions, mesh densities and time steps chosen were identical. One possible source is the accumulation of rounding errors, and another is that the truncation errors may have a slightly different form.

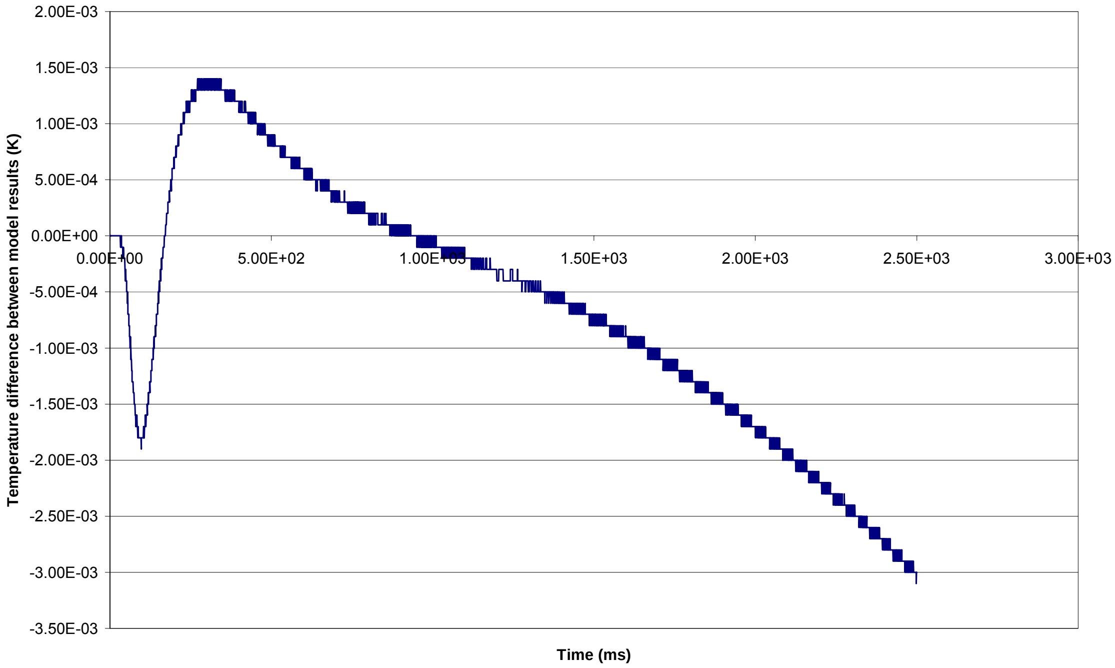

Figure 44: Difference between results of the full three-dimensional model and the axisymmetric equivalent.

Comparison with a simplified model is an internal consistency check since it tests the mathematical correctness of the implementation under simplifying assumptions.

8.2.7 Comparison with experimental data

The final test is to compare a full run of the model with analytic data gathered from an experiment with a material of known properties. The results of this test are shown in figure 45. The results were generated using the properties of Pyroceram (a ceramic with well-understood thermal behaviour) allowing for radiative cooling, and using a laser profile that was uniform in space but had a profile in time that was typical of that used in a real experiment. The l_2 error taken over 5000 points was 1.47 K. This is a good level of agreement, particularly since the experimental data was noisy and included the effects of the initial laser flash (shown in the figure as a sharp peak close to t = 0). When more data on non-uniform laser profiles is available this test will be repeated to assess the laser profile definition in the model.

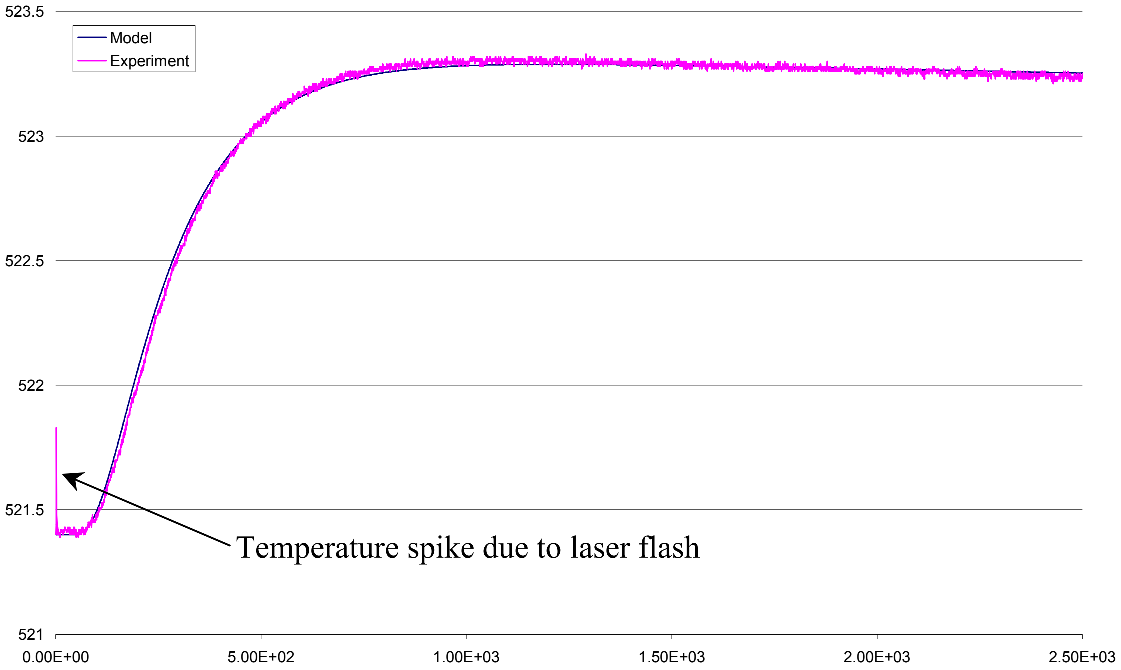

Figure 45: Comparison of experimental data and model results. The vertical axis has been removed to show the initial effects of the laser flash more clearly.

9 Summary of model validation in continuous modelling

The aim of this report has been to provide advice on the validation and testing of continuous models, particularly in relation to metrological applications. It was recognised that the validation philosophy that was developed in the SS*f*M-1 *Best Practice Guide to Discrete Model Validation* [1] can also be applied to continuous models. However, the use of continuous models in metrology often has a different motivation from the use of discrete models. Continuous models are most frequently used to obtain a deeper understanding of the process being modelled or to identify and quantify sources of systematic uncertainty in measurement processes, whereas discrete models are almost always applied to the understanding of experimental results and their associated uncertainties.

In the case of discrete models validation consists of internal and external consistency checks, but for continuous models an additional check of inter-model consistency is necessary, so that one can be satisfied that the discretisation method that has been chosen represents the problem of interest in a satisfactory manner. Thus, this report advocates a three-stage approach to the validation of continuous models. Is the model description adequate and mathematically correct? Is the discretisation method adequate? How do the results compare with other sources of information, including experimental results?

Section two of this report set out a detailed explanation of the mathematical notation that has been employed in the remainder of the report. The notation employed follows common practice in this branch of mathematical modelling, but is included both to assist readers who are not mathematicians and to avoid any possibility of confusion. The section also defines the use of norms to quantify the closeness of two sets of numbers and sets out definitions of consistency, stability and convergence.

Sources of error in continuous models were reviewed in section three. Five types of error were identified:

- modelling error (the failure of the chosen equation to represent the reality being modelled),
- space discretisation error (the error that arises from having to solve a continuous problem on a discrete spatial mesh),
- time discretisation error (similar to space discretisation error, the error in solving a continuous model at discrete time steps, with the added problem that time discretisation errors are accumulated at each time step),
- parameter errors (errors that arise from poorly chosen input parameters or from uncertainties associated with the values of the parameters in question), and
- linear algebra error (the errors generated during the solution of the system of equations generated by the model).

An overview of **validation strategies** was given in section four. The key recommendations are summarised below.

- The choice of validation methods depends strongly on the aim of the modelling process. Is one concerned simply to gain a qualitative understanding of the phenomenon of interest, or are the results to be used to quantify sources of uncertainty in a measurement process? In the second case the aim may be to develop the most comprehensive descriptive model possible, which implies an increased level of validation compared with the first case.

- Validation strategies should be part of the model development process, an aspect that is particularly important when a complex model is being constructed from simpler models or sub-models.
- Validation by comparison with experimental measurements can be problematic. Measurements may be available only for a limited part of the domain being modelled, and measurements may be unavailable or impossible to obtain in key parts of the model.
- Non-unique solutions may exist. For example, the physical problem may have a unique solution, but the mathematical formulation of the problem may have multiple solutions. Similarly, multiple physical solutions may co-exist. The most common case is hysteresis, where the response of a system to its load depends on its previous loading history as well as its current state. Although experimental repeatability may indicate that a unique stable solution exists, it is still possible that more than one stable solution can exist, but that its existence depends on an aspect of the experiment that may be too subtle to model.

Internal consistency checks on continuous models were discussed in section five. Topics covered were comparison with analytic solutions, comparison with equivalent models, and comparison with the results of other numerical methods. The main recommendations were:

- Comparison with an analytic solution is an ideal internal consistency check. It identifies computational errors and allows comparison with an alternative calculation method. The main disadvantage is that in many cases analytic solutions do not exist, which is usually why the numerical method was chosen originally. In addition, the model may not be able to reproduce the assumptions that were used in the calculation of the analytic solution.
- Analytic solutions may provide an effective check of finite element and finite difference software. If the software cannot accurately solve problems with known analytic solutions, it is unlikely to be able to produce reliable results for other cases.
- Comparisons with equivalent models can be useful. Two types of equivalent models were considered: simplifications of the model being validated and problems that are mathematically equivalent to the model. Clearly it is essential that the equivalent model has itself been validated.
- Simplified models include cases in which the assumption of symmetry may be used to reduce the problem size, the assumption of linearity of material properties (having the effect of reducing the complexity of the solution methods), and the assumption of constant properties (to reduce a dynamic problem to a static problem). The main caveat here is that the method of simplification obviously does not validate that part of the model that has been omitted for the purposes of the simplification.
- Different physical systems may be described by models that are mathematically equivalent. Examples include transport problems, where the diffusion of mass, heat or momentum may be described by the convection-diffusion equation, and the mechanical systems of masses, springs and dampers that obey the same

equations as electrical circuits formed of resistors, inductors and capacitors.

- It may be possible to use non-dimensional parameters to derive a single result that is applicable to a wide range of physical situations. A well-known example is the use of the Reynolds number in problems of steady-state viscous flow.
- Comparison with other numerical methods may also be useful, where such methods exist. Section six considered a specific example – the modelling of an acoustic field using two different boundary element formulations of the external Helmholtz problem. The approach requires that the model that is used for the validation check should have itself been validated, preferably by the use of both internal and external consistency checks. In addition, it will be necessary to take into account the accuracy of the validated model.

External consistency checks were reviewed in section six. Visual inspection of results was considered at length. Although such an inspection is a superficially attractive approach to model validation, especially as it allows the experience and knowledge of the experimentalist to be used in the validation process, it presents a number of difficulties. The main issues to be considered in the case of visual checks are:

- Often the most effective use of visual checks is to reject results that “look wrong” rather than to accept those that may “look right”.
- Initial visual checks should at least ensure that results are of the right order of magnitude and have the right sign. In addition, it is likely that results will vary smoothly, especially where the underlying equations represent diffusion processes of some kind.
- Where the modelling process itself has exploited symmetry or where results are expected to show symmetry, the results should be checked to ensure they exhibit the correct symmetrical behaviour.
- Visual inspection can be used to locate extreme values of parameters or to review the behaviour of the model at large distances from the main area of interest. Asymptotic behaviour can be checked this way.
- In most cases it is not possible to quantify the results of visual checks. Nevertheless, they should be documented and the results recorded, having the advantage that consistency can be maintained from one version of a model to the next.
- In the case of models with singularities – point loads and results at sharp, unfilleted corners – it may be possible to quantify the extent to which the effects of the singularity affect the remainder of the model. Section 6.1 of the report suggests a possible method for this quantification.

Additional advice in relation to external consistency checks includes the following:

- Many physical processes conserve quantities such as mass, momentum and energy. Several continuous modelling methods are derived from the application of conservation laws for these quantities. In cases where the physical process is expected to have a conservation law, the results should be checked to ensure that it is obeyed, particularly if the modelling method is derived from such a law.
- Specific problems in comparing modelling results with experimental data can arise where experimental results are contaminated by noise or interference. In

addition, it can be difficult to locate zeros in experimental processes, especially where the processes are time-dependent. Experimental uncertainty itself also contributes to the difficulty of comparing model results with experiment. Section 6.3 contains detailed advice on dealing with these difficulties.

- Some models generate more than one type of result. Finite element results, for example, may include displacements, stresses and strains. It is important to remember that validation of one set of results, such as the displacements, does not necessarily validate the stress results.
- Improved comparability between model and experiment may be observed if only a subset of the data is employed. For example, it may be possible to ignore differences at boundaries, corners and interfaces, or to ignore initial transients in time-dependent problems.
- Validation against experiment can be particularly difficult if it is not possible to obtain model parameters for aspects of the experimental set-up. An obvious example is the material properties for internal components of the equipment. It may be possible to use book values for material properties, but it is not usually certain that these values represent the real properties of the actual components of the experimental equipment.

Section seven took as its subject **inter-model consistency checks**, considering both a priori and a posteriori methods. These methods are concerned with the validity of the implementation of the model. For users of commercial and other “black box” software, it is likely that the aspect of the model over which they have most control is the mesh itself. Mesh-testing techniques can be very useful, as identifying and understanding errors introduced by the mesh may lead to cost-effective improvements to the model.

A priori methods: these methods seek to ensure that an approximate solution will converge to an exact solution as time and space discretisation steps tend to zero. Clearly, if a discretised model does not converge to an exact solution of the original differential equation, there is little point in trying to solve it. These methods often assume the existence of an exact solution and make assumptions about its behaviour. However, in general there will be little detailed information available about the exact solution, so that a priori methods do not normally produce quantifiable error estimates.

- Consistency checks are designed to ensure that the truncation error, which is generated over a single time step in a time-dependent problem, will tend to zero as the step size decreases. For finite difference schemes that have been developed using Taylor series expansions, proof of consistency may be relatively straightforward. In other cases, where meshes are less regular or of arbitrary shape, such checks may be more difficult.
- Time step stability analysis seeks to ensure that transient models do not use unstable time steps. The purpose is to avoid the exponential growth of round-off errors, which can often rapidly overshadow the true solution. The report contains detailed advice on this topic in section 7.1.2.
- Mesh convergence tests can often be as simple as re-running the problem with a denser mesh. This re-running can be carried out for both the spatial and the time mesh, with the risk of substantially increased model run times and computer memory requirements, especially in the case of reduced time steps.

A posteriori methods are concerned with estimating the error in the solution from the calculated discrete solution. Two main types exist: explicit methods, which use the computed solution and data from the problem itself to estimate the error, and implicit methods, which are based on the solution of a related problem, which is expressed in terms of an error function. Implicit methods are frequently time-consuming but may lead to a better estimate of the error than that provided by explicit methods, which tend to over-estimate error bounds. Sections 7.2.1 to 7.2.3 contain detailed advice on the mathematics of a posteriori tests for both the finite difference and finite element methods.

Finally, many of the recommendations summarised above were demonstrated on a test example described in section eight: a model of a thermal diffusivity experiment. This example demonstrated the combined use of the visual inspection of results, conservation of energy, consistency and stability tests, comparisons with an analytic solution and an equivalent model, and finally a comparison with experimental data. Although the tests did not employ every method advocated in this report, they are a representative cross-section of the methods that have been summarised above. Readers are therefore likely to find that they can easily adapt the approach advocated here to their own continuous modelling problems.

The systematic study of the application of continuous modelling methods in metrology is less advanced in the context of the Software Support for Metrology Programme than studies of discrete modelling. At present it may be too soon for best practice in continuous modelling to be defined. However, the authors hope that they have been able to provide readers with a basis of good practice in this increasingly important field and that future work may lead to towards the definition of best practice in the validation of continuous models for metrology.

10 References

[1] R. M. Barker and A. B. Forbes. Software Support for Metrology Best Practice Guide No 10: Discrete Model Validation. Available from http://www.npl.co.uk/ssfm/download/index.html#ssfmbpg10, March 2001.

[2] F. M. Hemez and S.W. Doebling. Model Validation And Uncertainty Quantification. Available from http://www.lanl.gov/projects/ncsd/pubs/IMAC-01_Paper3.pdf, October 2000.

[3] K. W. Morton and D. F. Mayers. Numerical solution of Partial Differential Equations. Cambridge University Press, 1994.

[4] B. P. Butler, M. G. Cox, A. B. Forbes, P. M. Harris and G. J. Lord. Model validation in the context of metrology: a survey. NPL Report CISE 19/99, February 1999.

[5] N. D. Fowkes and J. J. Mahony. An Introduction to Mathematical Modelling. John Wiley, 1994.

[6] H. A. Schenck. Improved integral formulation for acoustic radiation problems. J. Acoust. Soc. Am. 44(1), 41-58, 1968.

[7] A. J. Burton and G. F. Miller. The application of integral equation methods to the numerical solution of some exterior boundary-value problems. Proc. Roy. Soc. Lond. A 323, 201-210, 1971.

[8] O. C. Zienkewicz and R. L. Taylor. The Finite Element Method, Volume 1: Basic Formulation and Linear Problems. McGraw-Hill, 1994.

[9] M. Ainsworth and J. T. Oden. A Posteriori Error Estimation In Finite Element Analysis. Wiley Interscience, 2000.

[10] J. D. Lambert. Numerical Methods for Ordinary Differential Systems. Wiley, 1991.

[11] A. Umar, H. Abbas, A. Qadeer and D. K. Sehgal. Prediction of Error in Finite Element Results. Computers and Structures, v. 60 no. 3, 471-480, 1996.

[12] J. Mackerle. Error Estimates and Adaptive Finite Element Methods. A Bibliography 1990-2000. Engineering Computations, v.18, no. 5/6, 802-914, 2001.

[13] J. M. T. Thompson and H. B. Stuart. Nonlinear Dynamics and Chaos. Wiley, 2002.

[14] A. Stuart and A. R. Humphries. Dynamical Systems and Numerical Analysis. Cambridge University Press, 1998.

[15] E. Doedel et al. Auto97: Continuation and Bifurcation Software for ODEs. Technical Report, http://indy.cs.concordia.ca/auto/ for software details, 1997.

[16] ABAQUS Explicit v 5.5 User's Manual. Hibbitt, Karlsson and Sorensen, 1995.

[17] J. J. Rencis et al. A Posteriori Error Estimation for the Finite Element and Boundary Element Methods. Computers and Structures, v 37, no. 1, 103-117, 1990.

[18] A. L. Gurson. Continuum theory of ductile rupture by void nucleation and growth: Part I- Yield criteria and flow rules for porous ductile materials. J. Eng. Mater. Technol., v 99, 2-15, 1977.

[19] W. J. Parker, R. J. Jenkins, C. P. Butler and G. L. Abbott. Flash method of determining thermal diffusivity, heat capacity, and thermal conductivity. J. App. Phys.,

32(9), 1679-1684, September 1961.

[20] M. Križek and P. Neitaanmak. On Superconvergence Techniques. Acta Applicandae Mathematicae, v 9, 175-198, 1987.

[21] http://www.chem.ubc.ca/personnel/faculty/kast/Chem_205/Kinetics_week4.pdf

[22] T. J. Esward and L. Wright. Finite Element Optimisation For Problems In Acoustics. Proceedings of the National Measurement Conference 2001.

[23] BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, and OIML. Guide to the Expression of Uncertainty in Measurement. ISBN 92-67-10188-9, Second Edition, 1995.

[24] G. J. Lord and L. Wright. Uncertainty Evaluation in Continuous Modelling. NPL report No. CMSC 31/03, 2003.

[25] T. J. Esward and L. Wright. Guide to the use of finite element and finite difference software. NPL report No. 30/03, 2003.

[26] T. J. Esward, L. Wright et al. Testing Continuous Modelling Software. NPL report in preparation, 2004.

[27] P. G. Ciarlet and J. L. Lions (eds). Handbook of numerical analysis. Volume II: Finite element methods (part 1). ISBN 0-444-70365-9, 1991.

[28] M. Reader-Harris, C. D. Stewart, A. B. Forbes and G. J. Lord. Continuous Modelling In Metrology. NEL Report No. 058/2000 Rev 1, June 2000.

INDEX