Bayesian numerical methods for nonlinear partial differential equations

Junyang Wang[1] · Jon Cockayne[2] · Oksana Chkrebtii[3] · T. J. Sullivan[4] · Chris. J. Oates[1]

Abstract

The numerical solution of differential equations can be formulated as an inference problem to which formal statistical approaches can be applied. However, nonlinear partial differential equations (PDEs) pose substantial challenges from an inferential perspective, most notably the absence of explicit conditioning formula. This paper extends earlier work on linear PDEs to a general class of initial value problems specified by nonlinear PDEs, motivated by problems for which evaluations of the right-hand-side, initial conditions, or boundary conditions of the PDE have a high computational cost. The proposed method can be viewed as exact Bayesian inference under an approximate likelihood, which is based on discretisation of the nonlinear differential operator. Proof-of-concept experimental results demonstrate that meaningful probabilistic uncertainty quantification for the unknown solution of the PDE can be performed, while controlling the number of times the right-hand-side, initial and boundary conditions are evaluated. A suitable prior model for the solution of PDEs is identified using novel theoretical analysis of the sample path properties of Matérn processes, which may be of independent interest.

1 Introduction

Classical numerical methods for differential equations produce an approximation to the solution of the differential equation whose error (called *numerical error*) is uncertain in

✉ Chris. J. Oates
chris.oates@ncl.ac.uk

Junyang Wang
j.wang68@ncl.ac.uk

Jon Cockayne
jcockayne@turing.ac.uk

Oksana Chkrebtii
oksana@stat.osu.edu

T. J. Sullivan
t.j.sullivan@warwick.ac.uk

[1] School of Mathematics, Statistics and Physics, Newcastle University, Newcastle-Upon-Tyne NE1 7RU, UK

[2] Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK

[3] Department of Statistics, Ohio State University, 1958 Neil Avenue, 429 Cockins Hall, Columbus, OH 43210-1247, US

[4] Mathematics Institute and School of Engineering, The University of Warwick, Coventry CV4 7AL, UK

general. For non-adaptive numerical methods, such as finite difference methods, the extent of numerical error is often estimated by comparing approximations at different mesh sizes (Strikwerda 2004), while for adaptive numerical methods, such as certain finite element methods, the global tolerance that was given as an algorithmic input is used as a proxy for the size of the numerical error (Grätsch and Bathe 2005). On the other hand, probability theory provides a natural language in which uncertainty can be expressed and, since the solution of a differential equation is unknown, it is interesting to ask whether probability theory can be applied to quantify numerical uncertainty. This perspective, in which numerical tasks are cast as problems of statistical inference, is pursued in the field of *probabilistic numerics* (Larkin 1972; Diaconis 1988; O'Hagan 1992; Hennig et al. 2015; Oates and Sullivan 2019). A probabilistic numerical method (PNM) describes uncertainty in the quantity of interest by returning a probability distribution as its output. The Bayesian statistical framework is particularly natural in this context (Cockayne et al. 2019), since the output of a Bayesian PNM carries the formal semantics of *posterior uncertainty* and can be unambiguously interpreted. However, Wang et al. (2020) argued that a rigorous Bayesian treatment of ordinary differential equations (ODEs) may only be possible in a limited context

(namely, when the ODE admits a solvable Lie algebra). This motivates the development of approximate Bayesian PNMs, which aim to approximate the differential equation in such a way that exact Bayesian inference can be performed, motivated in particular by challenging numerical tasks for which numerical uncertainty cannot be neglected.

This paper focuses on partial differential equations (PDEs); in particular, we focus on PDEs whose governing equations must be evaluated pointwise at high computational cost. To date, no (approximate or exact) Bayesian PNM for the numerical solution of nonlinear PDEs has been proposed. However, the cases of nonlinear ODEs and linear PDEs have each been studied. In Chkrebtii et al. (2016) the authors constructed an approximate Bayesian PNM for the solution of initial value problems specified by either a nonlinear ODE or a linear PDE. The approach relied on conjugacy of Gaussian measures under linear operators to cycle between assimilating and simulating gradient field "data" in parallel. (Recall that the gradient field of an ODE can be a nonlinear function of the state variable, necessitating a linear approximation of the gradient field to obtain a conjugate Gaussian treatment.) The case of ODEs was discussed by Cockayne et al. (2019), who proposed a general computational strategy called *numerical disintegration* (similar to a sequential Monte Carlo method for rare event simulation) that can in principle be used to condition on nonlinear functionals of the unknown solution, such as provided by the gradient field, and thereby instantiate an *exact* Bayesian PNM. However, the authors emphasised that such methods are typically impractical. Tronarp et al. (2019, 2021) demonstrated how nonlinear filtering techniques can be used to obtain low-cost approximations to the PNM described by Cockayne et al. (2019) and applied their approach to approximate the solution of ODEs. Further work on approximate Bayesian PNM for ODEs includes Skilling (1992); Schober et al. (2014, 2019); Kersting and Hennig (2016); Kersting et al. (2020); Teymur et al. (2016, 2018); Chkrebtii and Campbell (2019); Krämer and Hennig (2020); Bosch et al. (2021). The focus of much of this earlier work was in the design of a PNM whose mean or mode coincides in some sense with a classical numerical method for ODEs (e.g. a Runge–Kutta method). For more detail on existing PNM for ODEs the reader is referred to the recent survey of Wang et al. (2020). Linear elliptic PDEs were considered by Särkkä (2011); Owhadi (2015); Cockayne et al. (2016) and Chkrebtii et al. (2016), and in this setting exact Gaussian conditioning can be performed. Chkrebtii et al. (2016) also presented a nonlinear PDE (Navier–Stokes), but a pseudo-spectral projection in Fourier space was applied at the outset to transform the PDE into a system of first-order ODEs — an approach that exploited the specific form of that PDE. Chen et al. (2021) extended Owhadi (2015) to nonlinear PDEs, focusing on *maximum a posteriori* estimation as opposed to uncertainty quantification for the solution of the PDE.

This paper presents the first (approximate) Bayesian PNM for numerical uncertainty quantification in the setting of nonlinear PDEs. Our strategy is based on local linearisation of the nonlinear differential operator, in order to perform conjugate Gaussian updating in an approximate Bayesian framework. Broadly speaking, our approach is a natural generalisation of the approach taken by Chkrebtii et al. (2016) for ODEs, but with local linearisation to address the additional challenges posed by nonlinear PDEs. The aim is to quantify numerical uncertainty with respect to the unknown solution of the PDE. An important point is that we consider only PDEs for which evaluation of either the right-hand side or the initial or boundary conditions is associated with a high computational cost; we do not aim to numerically solve PDEs for which a standard numerical method can readily be employed to drive numerical error to a negligible level, nor do we aim to compete with standard numerical methods in terms of CPU requirement. Such problems occur in diverse application areas, such as modelling of ice sheets, carbon and nitrogen cycles (Hurrell et al. 2013), species abundance and ecosystems (Fulton 2010), each in response to external forcing from a meteorological model, or in solving PDEs that themselves depend on the solution of an auxiliary PDE, which occur both when *operator splitting* methods are used (MacNamara and Strang 2016) and when *sensitivity equations*, expressing the rate of change of the solution of a PDE with respect to its parameters, are to be solved (Petzold et al. 2006; Cockayne and Duncan 2020). These applications provide strong motivation for PNM, since typically it will not be possible to obtain an accurate approximation to the solution of the PDE and the rich, probabilistic description of numerical uncertainty provided by a PNM can be directly useful (e.g. Oates et al. 2019,).

The remainder of the paper is structured as follows: In Sect. 2 the proposed method is presented. The choice of prior is driven by the mathematical considerations described in Sect. 3. A detailed experimental assessment is performed in Sect. 4. Concluding remarks are contained in Sect. 5.

2 Methods

In Sect. 2.1 we present the general form of the nonlinear PDE that we aim to solve using PNM. The use of finite differences for local linearisation is described in Sect. 2.2. Then, in Sect. 2.3 we present our proposed approximate Bayesian PNM, discussing how computations are performed and how the associated uncertainty is calibrated.

2.1 Set-up and notation

For a set $S \subseteq \mathbb{R}^d$, let $C^0(S)$ denote the vector space of continuous functions $c \colon S \to \mathbb{R}$. For two *multi-indices* $\alpha, \beta \in \mathbb{N}_0^d$,

we write $\alpha \leq \beta$ if $\alpha_i \leq \beta_i$ for each $i = 1, \ldots, d$. For a multi-index $\beta \in \mathbb{N}_0^d$, we let $|\beta| = \beta_1 + \cdots + \beta_d$ and let $C^\beta(S) \subseteq C^0(S)$ denote those functions c whose partial derivatives

$$\partial^\alpha c := \partial_{z_1}^{\alpha_1} \ldots \partial_{z_d}^{\alpha_d} c(z) := \frac{\partial^{|\alpha|} c(z)}{\partial z_1^{\alpha_1} \ldots \partial z_d^{\alpha_d}}, \qquad \alpha \leq \beta$$

exist and are continuous in S.

Let $T \in (0, \infty)$ and let Γ be an open and bounded set in $\mathbb{R}^d$, whose boundary is denoted $\partial\Gamma$. Let $\beta \in \mathbb{N}_0^d$ be a multi-index and consider a differential operator

$$D : C^\beta([0, T] \times \Gamma) \to C^0([0, T] \times \Gamma)$$

and the associated initial value problem with Dirichlet boundary conditions

$$\begin{aligned}
Du(t, x) &= f(t, x), & t &\in [0, T], \ x \in \Gamma \\
u(0, x) &= g(x), & x &\in \Gamma \\
u(t, x) &= h(t, x), & t &\in [0, T], \ x \in \partial\Gamma
\end{aligned} \qquad (2.1)$$

whose unique classical (i.e. strong) solution is denoted $u \in C^\beta([0, T] \times \Gamma)$ and is assumed to exist.[1] The task considered in this paper is to produce a probability distribution on $C^\beta([0, T] \times \Gamma)$ that (approximately) carries the semantics of Bayesian inference for u; i.e. we seek to develop an (approximate) Bayesian PNM for the numerical solution of Eq. (2.1) (Cockayne et al. 2019). In particular, we are motivated by the problems described in Sec. 1, for which evaluation of f, g and h are associated with a high computational cost. Such problems provide motivation for a careful quantification of uncertainty regarding the unknown solution u, since typically it will not be possible to obtain a sufficient number of evaluations of f, g and h in order for u to be precisely identified.

2.1.1 Why not emulation?

Given that the dominant computational cost is assumed to be evaluation of f, g and h, it is natural to ask whether the uncertainty regarding these functions can be quantified using a probabilistic model, such as an *emulator* (Kennedy and O'Hagan 2001). This would in principle provide a straightforward Monte Carlo solution to the problem of quantifying

uncertainty in the solution u of Eq. (2.1), where first one simulates an instance of f, g and h from the emulator and then applies a classical numerical method to solve Eq. (2.1) to high numerical precision. The problem with this approach is that construction of a defensible emulator is difficult; the functions f, g and h are coupled together by the nonlinear PDE in Eq. (2.1) and, for example, it cannot simultaneously hold that each of f, g and h are Gaussian processes. In fact, the challenge of ensuring that samples of f, g and h are consistent with the existence of a solution to Eq. (2.1) poses a challenge that is comparable with solving the PDE itself. This precludes a straight-forward emulation approach to Eq. (2.1) and motivates our focus on PNM in the remainder, where uncertainty is quantified in the solution space of Eq. (2.1).

2.2 Finite difference approximation of differential operators

If D is linear then the differential equation in Eq. (2.1) is said to be *linear* and one or more of the Bayesian PNM of Chkrebtii et al. (2016); Cockayne et al. (2016); Chkrebtii and Campbell (2019) may be applied (assuming any associated method-specific requirements are satisfied). If D is *nonlinear* then at most we can express $D = P + Q$, where P is linear and Q is nonlinear (naturally such representations are non-unique in general). For example, for Burgers' equation

$$Du = \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} - \varepsilon\frac{\partial^2 u}{\partial x^2} = 0$$

we have both $P = \partial_t - \varepsilon\partial_x^2$, $Q = u\partial_x$ and also the trivial $Q = D$, $P = 0$. In this paper we aim, given a decomposition of D in terms of P and Q, to adaptively approximate Q by a linear operator, in order that exact Gaussian conditioning formulae can be exploited. Although we do not prescribe how to select P and Q, one should bear in mind that we aim to construct a linear approximation of Q, meaning that a decomposition should be identified that renders Q as close to linear as possible, to improve the quality of the approximation. The effect of different selections for P and Q is investigated in Sect. 4.2.

To adaptively construct linear approximations to the nonlinear differential operator Q, we propose to exploit traditional finite difference formulae (Strikwerda 2004). Note that our conceptualisation of these approximations as linear *operators* for Gaussian conditioning is somewhat non-traditional. Define a time discretisation grid $\mathbf{t} = [t_0, t_1 \ldots t_{n-1}]$, where $0 = t_0 < t_1 < \cdots < t_{n-1} \leq T$ with the increment $\delta := t_i - t_{i-1}$ fixed. For concreteness, consider Burgers' equation with $P = \partial_t - \varepsilon\partial_x^2$, $Q = u\partial_x$. The following discussion is intended only to be informal. Suppose that the unknown solution $u(t_{i-1}, \cdot)$ at time t_{i-1} has been approximated to accuracy $\mathcal{O}(\delta)$ by $u_{i-1}(\cdot)$, as quantified by a norm $\|\cdot\|$ on

---

[1] The existence of a strong solution is a nontrivial assumption, since several PDEs admit only a weak solution; see Section 1.3.2 of Evans (1998) for definitions and background. A well-known class of classical numerical methods that also presuppose the existence of a strong solution are the radial basis function methods (Fornberg and Flyer 2015). In Sect. 4.2 we consider, empirically, the performance of the method developed in this paper when applied to a PDE for which a strong solution does not exist.

$C^\beta([0, T] \times \Gamma)$. Then we could adaptively build a linear approximation to Q at time t_i as

$$Q_i u(t_i, x) := u_{i-1}(x) \frac{\partial u}{\partial x}(t_i, x). \tag{2.2}$$

This provides an approximation $D_i = P + Q_i$ to the original differential operator D, at time t_i, with accuracy $\mathcal{O}(\delta)$. To achieve higher order accuracy, we can use higher order approximations of Q. For example, letting $\frac{\partial u}{\partial t}\big|_{i-1}(x)$ denote an approximation to $\frac{\partial u}{\partial t}(t_{i-1}, x)$, we could take

$$Q_i u(t_i, x) := \left[u_{i-1}(x) + \delta \left. \frac{\partial u}{\partial t}\right|_{i-1}(x) \right] \frac{\partial u}{\partial x}(t_i, x).$$

The only requirement that we impose on finite difference approximations is that Q_i uses (only) data that were gathered at earlier time points $t_{i-1}, t_{i-2}, \ldots$, analogous to *backward difference* formulae. This is to ensure that the approximations Q_i are well-defined before they are used in our method, which is described next.

Henceforth we assume that an appropriate representation $D = P + Q$ has been identified and an appropriate linear approximation to Q has been selected. The next section describes how probabilistic inference for u can then proceed.

2.3 Proposed approach

In this section we describe our proposed method. Recall that we assume there exists a unique $u \in C^\beta([0, T] \times \Gamma)$ for which Eq. (2.1) is satisfied. Since Eq. (2.1) represents an infinite number of constraints, it is not generally possible to recover u exactly with a finite computational budget. Our proposed method mirrors a general approach used to construct Bayesian PNM (Cockayne et al. 2019), in that we consider conditioning on only a finite number of the constraints in Eq. (2.1) and reporting the remaining uncertainty as our posterior. The case of nonlinear PDEs presents an additional challenge in that a subset of the constraints are nonlinear, and are therefore not amenable to exact Gaussian conditioning. To circumvent this issue, we condition on linear approximations to the constraints following the ideas developed in Sect. 2.2.

2.3.1 Prior distribution

The starting point of any Bayesian analysis is the elicitation of a suitable prior distribution. In our case, we require a prior that is supported on $C^\beta([0, T] \times \Gamma)$, since we *a priori* know that the solution u to Eq. (2.1) has this level of regularity. Our approach is rooted in Gaussian conditioning and thus the regularity of Gaussian process sample paths must be analysed. This analysis is somewhat technical and we therefore defer

the discussion of prior elicitation to Sect. 3. For the remainder of this section we assume a suitable Gaussian process prior $U \sim \mathcal{GP}(\mu, \Sigma)$ has been elicited. Here $\mu : [0, T] \times \Gamma \to \mathbb{R}$, $\mu(t, x) := \mathbb{E}[U(t, x)]$ is the *mean function* and $\Sigma : ([0, T] \times \Gamma) \times ([0, T] \times \Gamma) \to \mathbb{R}$, $\Sigma((t, x), (t', x')) := \mathbb{E}[(U(t, x) - \mu(t, x))(U(t', x') - \mu(t', x'))]$ is the *covariance function*; see Rasmussen and Williams (2006) for background. The random variable notation U serves to distinguish the true solution u of Eq. (2.1) from our probabilistic model for it. The specific choices of μ and Σ discussed in Sect. 3 have sufficient regularity for the subsequent derivations in this section to be well-defined.

2.3.2 Initialisation

At the outset we fix a time discretisation $\mathbf{t} = [t_0, t_1 \ldots t_{n-1}]$, where $0 = t_0 < t_1 < \cdots < t_{n-1} \le T$, and a spatial discretisation $\mathbf{x} = [x_1, x_2, \ldots, x_m] \in (\Gamma \cup \partial\Gamma)^m$ where the x_i are required to be distinct. It will sometimes be convenient to interpret $\mathbf{x}$ as a set $\{x_1, \ldots, x_n\}$; for instance we will write $\mathbf{x} \setminus \partial\Gamma$ to denote $\{x_1, \ldots, x_n\} \setminus \partial\Gamma$.

Our first task is to condition on (or *assimilate*) a finite number of constraints that encode the initial condition $u(0, x) = g(x)$, $x \in \Gamma$. For this purpose we use the spatial discretisation $\mathbf{x}$, and condition on the data $U(0, x) = g(x)$ at each $x \in \mathbf{x} \setminus \partial\Gamma$. (For example, if $\Gamma = [0, 1]$ and $0 = x_1 < x_2 < \cdots < x_{m-1} < x_m = 1$, then we condition on $U(0, x_i) = g(x_i)$ for $i = 2, \ldots, m - 1$. The two boundary locations $x_1, x_m \in \partial\Gamma$ are excluded since these constraints are assimilated as part of the boundary condition, which will shortly be discussed.) To perform conditioning, we use the following vectorised shorthand:

$$
\begin{aligned}
&\boldsymbol{v}_i := (t_i, \mathbf{x}) := [(t_i, x_1), (t_i, x_2), \ldots, (t_i, x_m)]^\top \\
&\quad \in ([0, T] \times \Gamma)^m \\
&U(\boldsymbol{v}_i) := [U(t_i, x_1), \ldots, U(t_i, x_m)]^\top \in \mathbb{R}^{m \times 1} \\
&g(\boldsymbol{v}_i) := [g(x_1), \ldots, g(x_m)]^\top \in \mathbb{R}^{m \times 1} \\
&\Sigma((t, x), \boldsymbol{v}_i) := [\Sigma((t, x), (t_i, x_1)), \ldots, \Sigma((t, x), (t_i, x_m))] \\
&\quad \in \mathbb{R}^{1 \times m} \\
&\Sigma(\boldsymbol{v}_i, (t, x)) := \Sigma((t, x), \boldsymbol{v}_i)^\top \in \mathbb{R}^{m \times 1} \\
&\Sigma(\boldsymbol{v}_i, \boldsymbol{v}_j) := \begin{bmatrix} \Sigma((t_i, x_1), (t_j, x_1)) & \cdots & \Sigma((t_i, x_1), (t_j, x_m)) \\ \vdots & & \vdots \\ \Sigma((t_i, x_m), (t_j, x_1)) & \cdots & \Sigma((t_i, x_m), (t_j, x_m)) \end{bmatrix} \\
&\quad \in \mathbb{R}^{m \times m}
\end{aligned}
$$

Then let $\boldsymbol{a}_0 := (t_0, \mathbf{x} \setminus \partial\Gamma)^\top$ denote the locations in $[0, T] \times \Gamma$ where the initial condition is to be assimilated. At $\boldsymbol{a}_0$ we have the *initial data* $\mathbf{y}^0 := g(\boldsymbol{a}_0)$. These initial data are assimilated into the Gaussian process model according to the standard conditioning formulae (eq. 2.19; Rasmussen and Williams 2006,)

$$U^0 := (U | U(\boldsymbol{a}_0) = \mathbf{y}^0) \sim \mathcal{GP}(\mu^0, \Sigma^0)$$

$$\mu^0(r) := \mu(r) + \Sigma(r, \boldsymbol{a}_0) \Sigma(\boldsymbol{a}_0, \boldsymbol{a}_0)^{-1}(g(\boldsymbol{a}_0) - \mu(\boldsymbol{a}_0))$$

$$\Sigma^0(r, s) := \Sigma(r, s) - \Sigma(r, \boldsymbol{a}_0) \Sigma(\boldsymbol{a}_0, \boldsymbol{a}_0)^{-1} \Sigma(\boldsymbol{a}_0, s)$$

where $r, s \in [0, T] \times \Gamma$.

2.3.3 Time stepping

Having assimilated the initial data, we now turn to the remaining constraints in Eq. (2.1). Following traditional time-stepping algorithms, we propose to proceed iteratively, beginning at time t_0 and then advancing to t_1, t_2, and ultimately to t_{n-1}. At each iteration i we aim to condition on a finite number of constraints that encode the boundary condition $u(t_i, x) = h(t_i, x), x \in \partial\Gamma$, and the differential equation itself $Du(t_i, x) = f(t_i, x), x \in \Gamma$. For this purpose we again use the spatial discretisation $\mathbf{x}$, and condition on the *boundary data* $U(t_i, x) = h(t_i, x)$ at each $x \in \mathbf{x} \cap \partial\Gamma$ and the *differential data* $DU(t_i, x) = f(t_i, x)$ at each $x \in \mathbf{x}$. Since D is nonlinear, there are no explicit formulae that can be used in general to assimilate the differential data, so instead we propose to condition on the approximate constraints $D_i U(t_i, x) = f(t_i, x), x \in \mathbf{x}$ where $D_i = P + Q_i$ is an adaptively defined linear approximation to D, which will be problem-specific and chosen in line with the principles outlined in Sect. 2.2.

For a univariate function such as μ and a linear operator L, we denote $\mu_L(r) = (L\mu)(r)$. For a bivariate function such as Σ, we denote $\Sigma_L(r, s) = L_r \Sigma(r, s)$, where L_r denotes the action of L on the r argument. In addition, we denote $\Sigma_{\bar{L}}(r, s) = L_s \Sigma(r, s)$ and we allow subscripts to be concatenated, such as $\Sigma_{L, L'} = (\Sigma_L)_{L'}$ for another linear operator L'.

Fix $i \in \{0, 1, \ldots, n-1\}$. Let $\boldsymbol{b}_i = (t_i, \mathbf{x} \cap \partial\Gamma)$ denote the locations in $[0, T] \times \partial\Gamma$ where the boundary conditions at time t_i are to be assimilated. At $\boldsymbol{b}_i$ we have boundary data $h(\boldsymbol{b}_i)$. Correspondingly, we have differential data $f(\boldsymbol{v}_i)$ and we concatenate all data at time i into a single vector $\mathbf{y}^i := [h(\boldsymbol{b}_i)^\top, f(\boldsymbol{v}_i)^\top]^\top$, so that $\mathbf{y}^i$ represents all the information on which (approximate) conditioning is to be performed. Upon assimilating these data we obtain

$$U^{i+1} := (U^i | [U(\boldsymbol{b}_i), D_i U(\boldsymbol{v}_i)] = \mathbf{y}^i)$$
$$\sim \mathcal{GP}(\mu^{i+1}, \Sigma^{i+1})$$

$$\mu^{i+1}(r) := \mu^i(r) +$$
$$[\Sigma^i(r, \boldsymbol{b}_i), \Sigma^i_{\bar{D}_i}(r, \boldsymbol{v}_i)] A_i^{-1} \begin{bmatrix} h(\boldsymbol{b}_i) - \mu^i(\boldsymbol{b}_i) \\ f(\boldsymbol{v}_i) - \mu^i_{D_i}(\boldsymbol{v}_i) \end{bmatrix}$$

$$\Sigma^{i+1}(r, s) := \Sigma^i(r, s) -$$
$$[\Sigma^i(r, \boldsymbol{b}_i), \Sigma^i_{\bar{D}_i}(r, \boldsymbol{v}_i)] A_i^{-1} \begin{bmatrix} \Sigma^i(\boldsymbol{b}_i, s) \\ \Sigma^i_{D_i}(\boldsymbol{v}_i, s) \end{bmatrix}$$

$$A_i := \begin{bmatrix} \Sigma^i(\boldsymbol{b}_i, \boldsymbol{b}_i) & \Sigma^i_{\bar{D}_i}(\boldsymbol{b}_i, \boldsymbol{v}_i) \\ \Sigma^i_{D_i}(\boldsymbol{v}_i, \boldsymbol{b}_i) & \Sigma^i_{D_i \bar{D}_i}(\boldsymbol{v}_i, \boldsymbol{v}_i) \end{bmatrix}$$

The result of performing n time steps of the algorithm just described is a Gaussian process $\mathcal{GP}(\mu^n, \Sigma^n)$, to which we associate the semantics of an (approximate) posterior in a Bayesian PNM for the solution of Eq. (2.1).

Remark 2.1 (Bayesian Interpretation) The Bayesian interpretation of $\mathcal{GP}(\mu^n, \Sigma^n)$ is reasonable since this distribution arises from the conditioning of the prior $\mathcal{GP}(\mu, \Sigma)$ on a finite number of constraints that are (approximately) satisfied by the solution u of Eq. (2.1). Indeed, from the self-consistency property of Bayesian inference (invariance to the order in which data are conditioned), the stochastic process U^n obtained above is identical to the distribution obtained when $U \sim \mathcal{GP}(\mu, K)$ is conditioned on the dataset

$$\begin{bmatrix} U(\boldsymbol{a}_0) \\ [U(\boldsymbol{b}_0), D_0 U(\boldsymbol{v}_0)]^\top \\ \vdots \\ [U(\boldsymbol{b})_{n-1}, D_n U(\boldsymbol{v}_{n-1})]^\top \end{bmatrix} = \begin{bmatrix} \mathbf{y}^0 \\ \mathbf{y}^1 \\ \vdots \\ \mathbf{y}^{n-1} \end{bmatrix}.$$

Remark 2.2 (Computational complexity) The computational cost of our algorithm is not competitive with that of a standard numerical method.[2] However, we are motivated by problems for which f, g and h are associated with a high computational cost, for which the auxiliary computation required to provide probabilistic uncertainty quantification is inconsequential. Thus we merely remark that the iterative algorithm we presented is gated by the inversion of the matrix A_i at the ith time step, the size of which is $O(m)$, independent of i, and therefore the complexity of predicting the final state $u(T, \cdot)$ of the PDE by performing n iterations of the above algorithm is $O(nm^3)$. For comparison, direct Gaussian conditioning on the information in Lemma 2.1 would incur a higher computational cost of $O(n^3m^3)$, but would provide the joint distribution over the solution $u(t, \cdot)$ at all times $t \in [0, T]$. Although we do not pursue it in this paper, in the latter case the grid structure present in $\mathbf{t}$ and $\mathbf{x}$ could be exploited to mitigate the $O(n^3m^3)$ cost; for example, a compactly supported covariance model Σ would reduce the cost by a constant factor (Gneiting 2002), or if the preconditions of Schäfer et al. (2021) are satisfied then their approach would reduce the cost to $O(nm \log(nm) \log^{d+1}(nm/\epsilon))$ at the expense of introducing an error of $O(\epsilon)$. See also the recent work of de Roos et al. (2021).

Remark 2.3 The posterior mean μ^{i+1} can be interpreted as a particular instance of a radial basis method (Fornberg and

---

[2] Technically, the computational complexity of our algorithm the same as that of a traditional numerical method that performs forward Euler increments in the temporal component and symmetric collocation in the spatial component (Fasshauer 1999; Cockayne et al. 2016). However such methods are rarely used, with one factor for this being the computational cost.

Flyer 2015), as a consequence of the representer theorem for kernel interpolants (Schölkopf et al. 2001). For brevity we do not explore this connection further, but we note that a similar connection was explored in detail in Cockayne et al. (2016).

2.3.4 Calibration of uncertainty

The principal advantage of PNM over classical numerical methods is that they provide probabilistic quantification of uncertainty, in our case expressed in the Bayesian framework, which can be integrated along with other sources of uncertainty to facilitate inferences and decision-making in a real-world context. In order for our posterior distribution to faithfully reflect the scale of uncertainty about the solution of Eq. (2.1), we must allow the hyper-parameters of the prior model to adapt to the dataset. However, we do not wish to sacrifice the sequential nature of our algorithm and thus we seek an approach to hyper-parameter estimation that operates in real-time as the algorithm is performed.

To achieve this we focus on a covariance model $\Sigma(\cdot, \cdot; \sigma)$ with a scalar hyper-parameter denoted $\sigma > 0$, which is assumed to satisfy $\Sigma(\cdot, \cdot; \sigma) = \sigma^2 \Sigma(\cdot, \cdot; 1)$. Such a σ is sometimes called a *scale* or *amplitude* hyper-parameter of the covariance model. From Lemma 2.1 it follows that σ directly controls the spread of the posterior and it is therefore essential that σ is estimated from data in order that the uncertainty reported by the posterior can be meaningful. To estimate σ, we propose to maximise the predictive likelihood of the "differential data" $f(\boldsymbol{v}_i)$, given the information collected up to iteration $i - 1$, for $i \in \{0, \ldots, n-1\}$, which can be considered as an *empirical Bayes* approach based on just those factors in the likelihood that correspond to the differential data. The reasons for focussing on the differential data (as opposed to also including the initial and boundary data) are twofold; first, the differential data constitutes the vast majority of the dataset, and second, this simplifies the computational implementation, described next.

At iteration i, the predictive likelihood for $U_{D_i}(\boldsymbol{v}_i)$ is $\mathcal{N}(\mu^i_{D_i}(\boldsymbol{v}_i), \Sigma^i(\boldsymbol{v}_i, \boldsymbol{v}_i; \sigma))$, and the observed differential data are $f(\boldsymbol{v}_i)$. Thus we select σ to maximise the full predictive likelihood of the differential data

$$\prod_{i=0}^{n-1} \mathcal{N}(f(\boldsymbol{v}_i); \mu^i_{D_i}(\boldsymbol{v}_i), \Sigma^i_{D_i}(\boldsymbol{v}_i, \boldsymbol{v}_i; \sigma)). \qquad (2.3)$$

Crucially, the linear operators D_i that we constructed do not depend on σ, and it is a standard property of Gaussian conditioning that $\Sigma^i_{D_i}(\boldsymbol{v}_i, \boldsymbol{v}_i; \sigma) = \sigma^2 \Sigma_{D_i}(\boldsymbol{v}_i, \boldsymbol{v}_i; 1)$. These facts permit a simple closed form expression for the maximiser $\hat{\sigma}$ of (2.3), namely

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=0}^{n-1} \left\| \Sigma^i_{D_i}(\boldsymbol{v}_i, \boldsymbol{v}_i; 1)^{-\frac{1}{2}} (f(\boldsymbol{v}_i) - \mu^i_{D_i}(\boldsymbol{v}_i)) \right\|^2 \qquad (2.4)$$

where $M^{-1/2}$ denotes an inverse matrix square root; $(M^{1/2})^2 = M$. Furthermore, it is clear from Eq. (2.4) that in practice one can simply run our proposed algorithm with the prior covariance model $\Sigma(\cdot, \cdot; 1)$ and then report the posterior covariance $\hat{\sigma}^2 \Sigma^{i+1}(\cdot, \cdot; 1)$, so that hyper-parameter estimation is performed in real-time without sacrificing the sequential nature of the algorithm.

Closed form expressions such as Eq. (2.4) are not typically available for other hyper-parameters that may be included in the covariance model, and we therefore assume in the sequel that any other hyper-parameters have been expert-elicited. This limits the applicability of our method to situations where some prior expert insight can be provided. However, we note that data-driven estimation of the amplitude parameter σ is able to compensate to a degree for mis-specification of other parameters in the covariance model.

2.3.5 Relation to earlier work

Here we summarise how the method just proposed relates to existing literature on Bayesian PNM and beyond.

The sequential updating procedure that we have proposed is similar to that of Chkrebtii et al. (2016) in the special case of a linear PDE. It is not identical in these circumstances though, for two reasons: First, Chkrebtii et al. (2016) incorporated the initial condition $u(0, x) = g(x)$, $x \in \Gamma$, into the prior model, whereas we explicitly conditioned on initial data $g(\boldsymbol{a}_0)$ during the initialisation step of the method. This direct encoding of the initial condition in Chkrebtii et al. (2016) relies on g being analytically tractable in order that a suitable prior can be derived by hand. Our treatment of g as a black-box function, which can (only) be point-wise evaluated, is therefore more general. Second, in Chkrebtii et al. (2016) the authors advocated the use of an explicit *measurement error* model, whereas our conditioning formula assume that the differential data $\boldsymbol{y}^i$ are exact measurements of U, as clarified in Lemma 2.1. For linear PDEs this assumption is correct, but it is an approximation in the case of a nonlinear PDE. Our decision not to employ a measurement error model here is due to the fact that the scale of the measurement error cannot easily be estimated in an online manner as part of a sequential algorithm, without further approximations being introduced.

To limit scope, the adaptive selection of the t_i or x_j was not considered, but we refer the reader to Chkrebtii and Campbell (2019) for an example of how this can be achieved using Bayesian PNM. Note, however, that adaptive selection of a time grid may be problematic when evaluation of either f or h is associated with a high computational cost, since

the possibility of taking many small time steps relinquishes control of the computational budget. For this reason, non-adaptive methods may be preferred in this context, since the run-time of the PNM can be provided up-front.

The choice of linearisation Q_i was left as an input to the proposed method, with some guidelines (only) provided in Sect. 2.2. This can be contrasted with recent work for ODEs in Tronarp et al. (2019, 2021); Bosch et al. (2021), where first-order Taylor series were used to automatically linearise a nonlinear gradient field. It would be possible to also consider the use of Taylor series methods for nonlinear PDEs. However, their use assumes that the gradient field is analytically tractable and can be differentiated, while in the present paper we are motivated by situations in which f is a black-box. The use of linearisations in PNM was also explored Chen et al. (2021), in the *maximum a posteriori* estimation context.

The combination of local linearisation and Gaussian process conditioning was also studied by Raissi et al. (2018), who considered initial value problems specified by PDEs, where the initial condition was random and the goal was to approximate the implied distribution over the solution space of the PDE. The authors observed that if the initial condition was a Gaussian process, then approximate conjugate Gaussian computation is possible when a finite difference approximation to the differential operator was employed. This provided a one-pass, cost-efficient alternative to the Monte Carlo approach of repeatedly sampling an initial condition and then applying a classical numerical method. Our work bears a superficial similarity to Raissi et al. (2018) and related work on *physics-informed* Gaussian process regression (e.g. Wheeler et al. 2014; Wang and Berger 2016; Jidling et al. 2017; Chen et al. 2020,), in that finite difference approximations enable approximate Gaussian conditioning to be performed. However, these authors are addressing a *fundamentally different* problem to that addressed in the present paper; we aim to quantify numerical uncertainty for a single (i.e. non-random) PDE. Accordingly, in this paper we emphasise issues that are critical to the performance of PNM, such as ensuring that the posterior is supported on a set of functions whose regularity matches that known to be possessed by the solution of the PDE Sect. (3), and explicitly assessing the quality of the credible sets provided by our PNM Sect. (4).

3 Prior construction

This section is dedicated to presenting a prior construction that ensures samples generated from the prior are elements of $C^\beta([0, T] \times \Gamma)$, the set in which a solution to Eq. (2.1) is sought. First, in Sect. 3.1 we introduce the technical notions of *sample continuity* and *sample differentiability*, clarify-

ing what properties of the prior are required to hold. These sample-path properties are distinct from mean-square properties, the latter being more commonly studied. Then, in Sect. 3.2 we formally prove that the required properties hold for a particular Matérn tensor product, which we then advocate as a default choice for our PNM. These results may also be of independent interest.

3.1 Mathematical properties for the prior

This paper is concerned with the strong solution of Eq. (2.1), which is an element of $C^\beta([0, T] \times \Gamma)$. It is therefore logical to construct a prior distribution whose samples also belong to this set. In particular, if the true solution has β_i derivatives in the variable z_i (for instance because the PDE features a term $\partial_{z_i}^{\beta_i} u$, where we have set $z = (t, x)$), it would be appropriate to construct a prior (and hence a posterior) whose samples also have β_i derivatives in the variable z_i.

To make this discussion precise, we make explicit a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and recall the fundamental definitions of sample continuity and sample differentiability for a random field $X : I \times \Omega \to \mathbb{R}$ defined on an open, pathwise-connected set $I \subseteq \mathbb{R}^d$ (i.e. I is an interval when $d = 1$):

Definition 3.1 *(Sample Continuity) X is said to be *sample continuous* if, for $\mathbb{P}$-almost all $\omega \in \Omega$, the sample path $X(\cdot, \omega)$ is continuous (everywhere) in I.*

Definition 3.2 *(Sample Differentiability) Consider a sequence $v^1, \ldots, v^p \in \mathbb{R}^d$ of directions and let $v = (v^1, \ldots, v^p)$. Then X is said to be *sample partial differentiable in the sequence of directions* v if for $\mathbb{P}$-almost all $\omega \in \Omega$, the following limit exists for all $z \in I$*

$$\mathcal{D}^p X(z, v, \omega) = \lim_{h_1 \to 0} \cdots \lim_{h_p \to 0} \frac{\Delta^p X(z, v, h, \omega)}{\prod_{i=1}^p h_i} < \infty$$

where $\Delta^p X(z, v, h, \omega) :=$

$$\sum_{r \in \{0,1\}^p} (-1)^{p - \sum_{i=1}^p r_i} X\left(z + \sum_{i=1}^p r_i h_i v^i, \omega\right).$$

The limits above are taken sequentially from left to right. In the discussions that follow, we take v^i from $\{e_1, e_2, \ldots, e_d\}$, the standard Cartesian unit basis vectors of $\mathbb{R}^d$, in which case the usual partial derivatives are retrieved, and we use the shorthand $\mathcal{D}^p X(z, v, \omega) = \partial^\alpha X(z, \omega)$ to denote sample partial derivatives, where $\alpha = (\alpha_1, \ldots, \alpha_p)$, $|\alpha| = p$, and α_i denotes the number of times the variable z_i is differentiated.

A similar property, which is more easily studied than sample continuity (resp. sample differentiability), is mean-square continuity (resp. mean-square differentiability). This property is recalled next, since we will make use of mean-square

properties en route to establishing sample path properties in Sect. 3.2.

Definition 3.3 (*Mean-Square Continuity*) X is said to be *mean-square continuous at $z \in I$* if

$$\mathbb{E}\left[X(z,\omega)^2\right] < \infty, \quad \lim_{z' \to z} \mathbb{E}\left[(X(z',\omega) - X(z,\omega))^2\right] = 0.$$

Definition 3.4 (*Mean-Square Differentiability*) Let $v^1, \ldots, v^p \in \mathbb{R}^d$ be a sequence of directions and $v = (v^1, \ldots, v^p)$. Then X is said to be *mean-square partial differentiable at $z \in I$ in the sequence of directions v* if there exists a finite random field $\omega \mapsto \mathcal{D}^p_{\mathrm{MS}} X(z, v, \omega)$ such that

$$\lim_{h_1 \to 0} \cdots \lim_{h_p \to 0} \mathbb{E}\left[\left(\frac{\Delta^p X(z, v, h, \omega)}{\prod_{i=1}^p h_i} - \mathcal{D}^p_{\mathrm{MS}} X(z, v, \omega)\right)^2\right] = 0$$

is well-defined.

For a mean-square differentiable Gaussian processes X, with mean function $\mu \in C^\alpha(I)$ and covariance function $\Sigma \in C^{(\alpha,\alpha)}(I \times I)$, one has

$$\partial^\alpha_{\mathrm{MS}} X \sim \mathcal{GP}(\partial^\alpha \mu, \partial^\alpha \bar{\partial}^\alpha \Sigma)$$

where we use the shorthand $\mathcal{D}^p_{\mathrm{MS}} X(z, v, \omega) = \partial^\alpha_{\mathrm{MS}} X(z, \omega)$ to denote mean-square partial derivatives, where again the v^i are unit vectors parallel to the coordinate axes, $|\alpha| = p$, and α_i denotes the number of times the variable z_i is differentiated.[3] See (Stein, 1999, Section 2.6). If X is mean-square continuous (resp. mean-square differentiable for all $\alpha \leq \beta$) at all $z \in I$, then we say simply that X is *mean-square continuous* (resp. *order β mean-square differentiable*). In contrast to sample path properties, mean-square properties are often straight-forward to establish. In particular, if X is weakly stationary with autocovariance function $\Sigma(z) = \Sigma(z, 0)$, then

$$\mathbb{E}\left[(X(z,\omega) - X(z',\omega))^2\right] = 2(\Sigma(0) - \Sigma(z - z')), \quad (3.1)$$

meaning that X is mean-square continuous whenever its autocovariance function Σ is continuous at 0 (Stein 1999, Section 2.4).

3.2 Matérn tensor product

Our aim in this section is to establish sample path properties for a particular choice of prior, namely a Gaussian process

with tensor product Matérn covariance, to ensure that prior (and posterior) samples are contained in $C^\beta([0, T] \times \Gamma)$. Surprisingly, we are unable to find explicit results in the literature for the sample path properties of commonly used covariance models; this is likely due to the comparative technical difficulty in establishing sample path properties compared to mean-square properties. Our aim in this section is, first, to furnish a gap in the literature by rigorously establishing the sample differentiability properties of Gaussian processes defined by the Matérn covariance function and, second, to put forward a default prior for our PNM that takes values in $C^\beta([0, T] \times \Gamma)$.

Definition 3.5 (*Matérn Covariance*) Let $\nu = p + \frac{1}{2}$ where $p \in \mathbb{N}$ and let $\rho \in (0, \infty)$. The Matérn covariance function is defined, for $z, z' \in \mathbb{R}$, as

$$K_\nu(z, z') = K_\nu(z - z')$$
$$= \sigma^2 \exp\left(-\frac{|z-z'|}{\rho}\right) \frac{p!}{(2p)!} \sum_{k=0}^p \frac{(2p-k)!}{(p-k)!k!} \left(\frac{2}{\rho}\right)^k |z - z'|^k. \quad (3.2)$$

The proof of the following result can be found in 1.

Proposition 3.1 *Let $I \subseteq \mathbb{R}$ be an open set and let $\mu \in C^p(I)$. Then any process $X \sim \mathcal{GP}(\mu, K_\nu)$, with K_ν as in Eq. (3.2) with $\nu = p + \frac{1}{2}$, is order p mean-square differentiable. Furthermore, $\partial^p_{\mathrm{MS}} X$ is mean-square continuous.*

Following a general approach outlined in Potthoff (2010), and focussing initially on the univariate case, our first step toward establishing sample differentiability is to establish sample continuity of the mean-square derivatives. Recall that, for two stochastic processes $X, \tilde{X}$ on a domain I, we say $\tilde{X}$ is a *modification* of X if, for every $z \in I$, $\mathbb{P}(X(z,\omega) = \tilde{X}(z,\omega)) = 1$. A modification of a stochastic process does not change its mean square properties, but sample path properties need not be invariant to modification.[4] For Gaussian processes, which are characterised up to modifications by their finite dimensional distributions, it is standard practice to work with continuous modifications when they exist (see for example Dudley 1967; Marcus and Shepp 1972,). The proof of the following result can be found in 1.

Proposition 3.2 *Let X be as in 3.1. Then $\partial^i_{\mathrm{MS}} X$ has a modification that is sample continuous for all $0 \leq i \leq p$.*

---

[3] The shorthand notation suppresses the order in which derivatives are taken, and can therefore only be applied in situations where partial derivatives are continuous, to ensure that their order can be interchanged without affecting the result. In the following, the notation $\partial^\alpha_{\mathrm{MS}} X$ is used only for Gaussian processes with $\partial^\alpha \bar{\partial}^\alpha \Sigma \in C(I \times I)$.

[4] To build intuition into the role of modifications, let $X : [0, 1] \times \Omega \to \mathbb{R}$ be a sample continuous stochastic process and consider the process $\tilde{X}(z,\omega) := X(z,\omega) + 1[z = Z(\omega)]$ where $Z \sim \mathcal{U}(0, 1)$, independent of X. Then $\tilde{X}$ is a modification of X whose finite-dimensional distributions (and hence mean square properties) are identical to those of X, but $\tilde{X}$ is almost surely *not* sample continuous. In such circumstances it is convenient (and standard practice) to work with the sample continuous process, X, as opposed to $\tilde{X}$.

The second step is to leverage a fundamental result on the sample path properties of stochastic processes; Theorem 3.2 of Potthoff ([2010]):

Theorem 3.1 *Let $I \subseteq \mathbb{R}^d$ be an open, pathwise connected set, and consider a random field $X : I \times \Omega \to \mathbb{R}$ such that $\mathbb{E}[X(z, \omega)^2] < \infty$ for all $z \in I$. Suppose X is first order mean-square differentiable, with mean-square partial derivatives $\mathcal{D}^1_{\mathrm{MS}} X(\cdot, e_k, \omega)$, $1 \leq k \leq d$, themselves being mean-square continuous and having modifications that are sample continuous. Then X has a modification $\tilde{X}$ that is first order sample partial differentiable, with partial derivatives $\mathcal{D}^1 \tilde{X}(\cdot, e_k, \omega)$, $1 \leq k \leq d$, themselves being sample continuous and satisfying, almost surely, $\mathcal{D}^1 \tilde{X}(\cdot, e_k, \omega) = \mathcal{D}^1_{\mathrm{MS}} X(\cdot, e_k, \omega)$, $1 \leq k \leq d$.*

Since continuity of partial derivatives implies differentiability, the conclusion of 3.1 implies that X is first order sample differentiable .

Iterative application of 3.1 to higher order derivatives provides the following, whose proof can be found in Section 6.3.

Corollary 3.1 *Fix $p \in \mathbb{N}$. Let $I \subseteq \mathbb{R}^d$ be an open, pathwise connected set and consider $X \sim \mathcal{GP}(\mu, \Sigma)$ with $\mu \in C^p(I)$ and $\Sigma \in C^{(p,p)}(I \times I)$, so that X has mean-square partial derivatives $\partial^\beta_{\mathrm{MS}} X$, $\beta \in \mathbb{N}_0^d$, $|\beta| \leq p$. Suppose $\partial^\beta_{\mathrm{MS}} X$ is mean-square continuous and sample continuous for all $|\beta| \leq p$. Then X has continuous sample partial derivatives $\partial^\beta X$, and they satisfy $\partial^\beta X = \partial^\beta_{\mathrm{MS}} X$ almost surely, for all $|\beta| \leq p$.*

This provides a strategy to establish sample properties of Matérn processes, such as the following:

Corollary 3.2 *Let $I \subseteq \mathbb{R}$ be an open interval and let $X \sim \mathcal{GP}(\mu, K_\nu)$, with $\mu \in C^p(I)$ and K_ν as in Eq. (3.2). Then there exists a modification $\tilde{X}$ of X such that $\mathbb{P}(\tilde{X} \in C^p(I)) = 1$.*

Proof By 3.1, X is order p mean-square differentiable and $\partial^i_{\mathrm{MS}} X$ is mean-square continuous for $0 \leq i \leq p$. By 3.2, we may work with a modification $\tilde{X}$ of X such that $\partial^i_{\mathrm{MS}} \tilde{X} (= \partial^i_{\mathrm{MS}} X)$ is sample continuous for each $0 \leq i \leq p$. One can directly verify that $K_\nu \in C^{(p,p)}(I \times I)$; see the calculations in 1. The result then follows from 3.1. □

Equation 3.2 is stronger than existing results in the literature, the most relevant of which is (Scheuerer, [2010], Theorem 5), who showed that samples from $\mathcal{GP}(\mu, K_{\nu+\epsilon})$ are $C^p(I)$ for any $\epsilon > 0$.

Finally, we present a multivariate version of the previous result, which will be exploited in the experiments that we perform in Sect. 4. Importantly, we allow for different smoothness in the different variables, which is necessary to properly capture the regularity of solutions to PDEs. The proof of this result is contained in 1.

Theorem 3.2 *Let $I = (a_1, b_1) \times \cdots \times (a_d, b_d)$ be a bounded hyper-rectangle in $\mathbb{R}^d$. Fix $\beta \in \mathbb{N}_0^d$. Let $\mu \in C^\beta(I)$ be bounded[5] in I, and consider a covariance function $\Sigma : I \times I \to \mathbb{R}$ of the form*

$$\Sigma(z, z') = \prod_{i=1}^{d} K_{\nu_i}(z_i - z_i')$$

where $z = (z_1, z_2, \ldots, z_d)$, $z' = (z_1', z_2', \ldots, z_d')$ and $\nu_i = \beta_i + \frac{1}{2}$ for each $i = 1, \ldots, d$. Then a Gaussian process of the form $X \sim \mathcal{GP}(\mu, \Sigma)$ has a modification $\tilde{X}$ that satisfies $\mathbb{P}(\tilde{X} \in C^\beta(I)) = 1$.

4 Experimental assessment

In this section, proof-of-concept numerical studies of three different initial value problems are presented. The first and simplest case is a homogeneous Burger's equation, a PDE with one nonlinear term and a solution that is known to be smooth. The second case is a porous medium equation, with two nonlinear terms appearing in the PDE and a solution that is known to be piecewise smooth, so that a classical solution does not exist and our modelling assumptions are violated. The third case returns to Burger's equation but now with forcing, to simulate a scenario where the right hand side f is a black box function that may be evaluated at a high computational cost. All three experiments are *synthetic*, in the sense that the functions f, g and h, which in our motivating task are considered to be black boxes associated with a high computational cost, are in actual fact simple analytic expressions, enabling a thorough empirical assessment to be performed.

In order to assess the empirical performance of our algorithm, two distinct performance measures were employed. The first of these aims to assess the accuracy of the posterior mean, which is analogous to how classical numerical methods are assessed. For this purpose the L^∞ error was considered:

$$E_\infty := \sup_{t \in [0,T], x \in \Gamma} \left| \mu^n(t, x) - u(t, x) \right| \tag{4.1}$$

In practice the value of Eq. (4.1) is approximated by taking the maximum over the grid $\mathbf{t} \times \mathbf{x}$ on which the data $\mathbf{y}^0, \ldots, \mathbf{y}^{n-1}$ were obtained. Accuracy that is comparable to a classical numerical method is of course desirable, but it is not our goal to compete with classical numerical methods in terms of L^∞ error. The second statistic that we consider assesses whether the distributional output from our PNM is *calibrated*, in the

---

[5] The requirement that μ be bounded, which did not appear in the corresponding univariate result (3.2) is likely an artefact of our proof strategy, rather than a necessary condition for 3.2 to hold.

sense that the scale of the Gaussian posterior is comparable with the difference between the posterior mean μ^n and the true solution u of Eq. (2.1):

$$
Z := \sup_{t \in [0,T], x \in \Gamma} \frac{|\mu^n(t,x) - u(t,x)|}{\hat{\sigma} \, \Sigma^n(t,x)^{1/2}} \tag{4.2}
$$

This performance measure will be called a *Z-score*, in analogy with traditional terminology from statistics. For the purpose of this exploratory work, values of Z that are orders of magnitude smaller than 1 are interpreted as indicating that the distributional output from the PNM is under-confident, while values that are orders of magnitude greater than 1 indicate that the PNM is over-confident. A PNM that is neither under nor over confident is said to be *calibrated* (precise definitions of the term "calibrated" can be found in Karvonen et al. 2020; Cockayne et al. 2021, but the results we present are straight-forward to interpret using the informal approach just described). Our goal in this work is to develop an approximately Bayesian PNM for nonlinear PDEs that is both accurate and calibrated. Again, in practice the supremum in Eq. (4.2) is approximated by the maximum over the **t** × **x** grid.

For all experiments below, we consider uniform temporal and spatial grids of respective sizes $n = 2^i + 1$, $m = 2^j + 1$, where $i, j \in \{2, 3, 4, 5, 6, 7\}$. This ensures that the grid points at which data are obtained are strictly nested as either the temporal exponent i or the spatial exponent j are increased. The prior mean $\mu(t,x) = 0$ for all $t \in [0,T]$, $x \in \Gamma$, will be used throughout.

4.1 Homogeneous Burger's equation

Our first example is the homogeneous *Burger's equation*

$$
\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} - \alpha \frac{\partial^2 u}{\partial x^2} = 0, \qquad t \in [0,T], \ x \in [0,L]
$$

with initial and boundary conditions

$$
u(0,x) = 2\alpha \left(\frac{ak\sin(kx)}{b + a\cos(kx)} \right), \qquad x \in [0,L]
$$

$$
u(t,0) = u(t,2\pi) = 0, \qquad t \in [0,T]
$$

and, for our experiments, $\alpha = 0.02$, $a = 1$, $b = 2$, $k = 1$, $T = 30$ and $L = 2\pi$. These initial and boundary conditions were chosen because they permit a closed-form solution

$$
u(t,x) = 2\alpha \left(\frac{ak\exp(-\alpha k^2 t)\sin(kx)}{b + a\exp(-\alpha k^2 t)\cos(kx)} \right) \tag{4.3}
$$

that can be used as a ground truth for our assessment.

To linearise the differential operator Burger's equation we consider approximations of the form in Eq. (2.2), i.e.

$$
Q_i u(t_i, x) := u_{i-1}(x) \frac{\partial u}{\partial x}(t_i, x)
$$

where $u_{i-1}(x)$ was taken equal to the predictive mean $\mu^{i-1}(t_i, x)$ arising from the Gaussian process approximation U^{i-1}.

4.1.1 Default prior

Burger's equation has a first-order temporal derivative term and a second-order spatial derivative term, so following the discussion in Sect. 3 we consider as a default a Gaussian process prior with covariance function Σ that is a product between a Matérn 3/2 kernel $K_{3/2}(t, t')$ for the temporal component, and a Matérn 5/2 kernel $K_{5/2}(x, x')$ for the spatial component:

$$
\Sigma((t,x), (t',x')) = K_{3/2}(t, t'; \rho_1, \sigma_1) \\
K_{5/2}(x, x'; \rho_2, \sigma_2) \tag{4.4}
$$

The notation in Eq. (4.4) makes explicit the dependence on the amplitude hyper-parameters σ_1, σ_2 and the length-scale hyper-parameters ρ_1, ρ_2; note that only the product $\sigma := \sigma_1 \sigma_2$ of the two amplitide parameters is required to be specified. For the experiments below σ was estimated as per Eq. (2.4), while the length-scale parameters were fixed at values $\rho_1 = 6$, $\rho_2 = 3$ (not optimised; these were selected based on a *post-hoc* visual check of the credible sets in Fig. 1). From 3.2, this construction ensures that the prior is supported on $C^{(1,2)}([0,T] \times [0,L])$. Typical output from our PNM equipped with the default prior is presented in Fig. 1.

4.1.2 An alternative prior

The Matérn covariance models assume only the minimal amount of smoothness required for the PDE to be well-defined. However, in this assessment the ground truth u is available Eq. (4.3) and is seen to be infinitely differentiable in $(0,T] \times [0, 2\pi]$. It is therefore interesting to explore whether a prior that encodes additional smoothness can improve on the default prior in Eq. (4.4). A prototypical example of such a prior is

$$
\Sigma((t,x), (t',x')) = C(t, t'; \rho_3, \sigma_3) C(x, x'; \rho_4, \sigma_4)
$$

where

$$
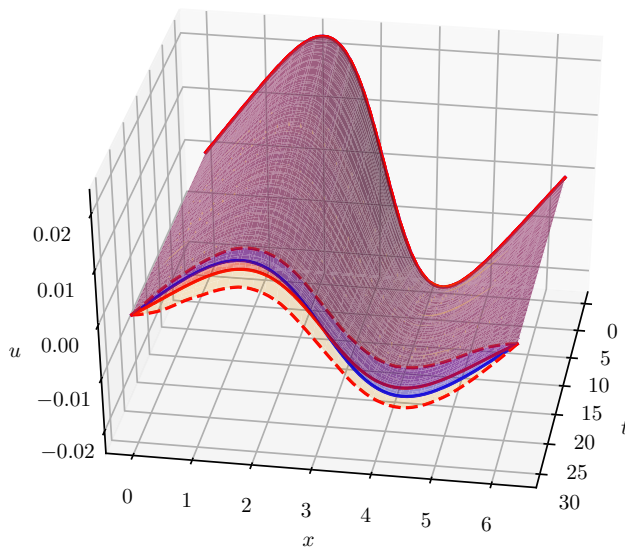C(z, z'; \rho, \sigma) := \sigma \left(1 + \frac{(z - z')^2}{\rho^2} \right)^{-1}
$$

Fig. 1 Homogeneous Burger's equation: For each point (t, x) in the domain we plot: the analytic solution $u(t, x)$ (blue), the posterior mean $\mu^n(t, x)$ (red) from the proposed probabilistic numerical method, and 0.025 and 0.975 quantiles of the posterior distribution at each point (orange). Here the default prior was used, with a spatial grid of size $m = 65$ and a temporal grid of size $n = 65$. (Color figure online)

is the *rational quadratic* covariance model. For the experiments below σ was estimated as per Eq. (2.4), while the length-scale parameters were fixed at values $\rho_3 = \sqrt{3}$, $\rho_4 = \sqrt{3}$.

4.1.3 Results

The error E_∞ was computed at 36 combinations of temporal and spatial grid sizes (n, m) and results for the default prior are displayed in the top left of Fig. 2. It can be seen that the error E_∞ is mostly determined, in this example, by the finite length n of the temporal grid rather than the length m of the spatial grid. The slope of the curves in Fig. 2 is consistent with a convergence rate of $O(n^{-1})$ for the error E_∞ when spatial discretisation is neglected. The Z-scores associated with the default prior (top right of Fig. 2) appear to be bounded as (n, m) are increased, tending toward 0 but taking values of order 1 for all regimes, except for the smallest value ($m = 5$) of the spatial grid. This provides evidence that our proposed PNM, equipped with the default prior, is either calibrated or slightly under-confident but, crucially from a statistical perspective, it is not over-confident.

Equivalent results for the alternative covariance model are presented for the error E_∞ in the bottom left of Fig. 2 and for the Z-score in the bottom right of Fig. 2. Here the error E_∞ is again gated by the size n of the temporal grid and decreases at a faster rate compared to when the default prior was used. This is perhaps expected because the rational quadratic covariance model reflects the true smoothness of the solu-

tion better than the Matérn model. However, the Z-scores associated with the alternative covariance model are considerably higher, appearing to grow rapidly as $n \to \infty$ with m fixed. This suggests that the alternative covariance model is inappropriate, causing our PNM to be over-confident. We speculate this may be because u does not belong to the support of the rational quadratic covariance model, but note that the support of a Gaussian process can be difficult to characterise (Karvonen 2021). These results support our proposed strategy for prior selection in Sect. 3.

4.2 Porous medium equation

Our second example is the *porous medium equation*

$$\frac{\partial u}{\partial t} - \frac{\partial^2 (u^k)}{\partial x^2} = 0, \ t \in [t_0, t_0 + T], \ x \in \left[-\frac{L}{2}, \frac{L}{2}\right], \ (4.5)$$

which is more challenging compared to Burger's equation because the solution is only piecewise smooth, meaning that a strong solution does not exist and our modelling assumptions are violated. Furthermore, there are two distinct nonlinearities in the differential operator, allowing us to explore the impact of the choice of linearisation on the performance of the PNM. For our experiment we fix $k = 2$, so that the porous medium equation becomes

$$\frac{\partial u}{\partial t} - 2 \left(\frac{\partial u}{\partial x}\right)^2 - 2u \frac{\partial^2 u}{\partial x^2} = 0,$$
$$t \in [t_0, t_0 + T], \ x \in \left[-\frac{L}{2}, \frac{L}{2}\right],$$

and we consider the initial and boundary conditions

$$u(t_0, x) = t_0^{-1/3} \max \left(0, 1 - \frac{x^2}{12 t_0^{2/3}}\right), \ x \in \left[-\frac{L}{2}, \frac{L}{2}\right]$$
$$u(t, -L/2) = u(t, L/2) = 0, \ t \in [t_0, t_0 + T]$$

with $t_0 = 2$, $T = 8$, $L/2 = 10$. These initial and boundary conditions were chosen because they permit a (unique) closed-form solution, due to Barenblatt (1952):

$$u(t, x) = \max \left(0, \frac{1}{t^{1/3}} \left(1 - \frac{1}{12} \frac{x^2}{t^{2/3}}\right)\right)$$

The solution is therefore only piecewise smooth, with discontinuous first derivatives at $x^2 = 12 t^{2/3}$, which are inside of the domain $[-L/2, L/2]$ for all $t \in [t_0, t_0 + T]$.

4.2.1 Prior

Henceforth we consider the default prior advocated in Sects. 3 and 4.1, with amplitude σ estimated using maximum like-
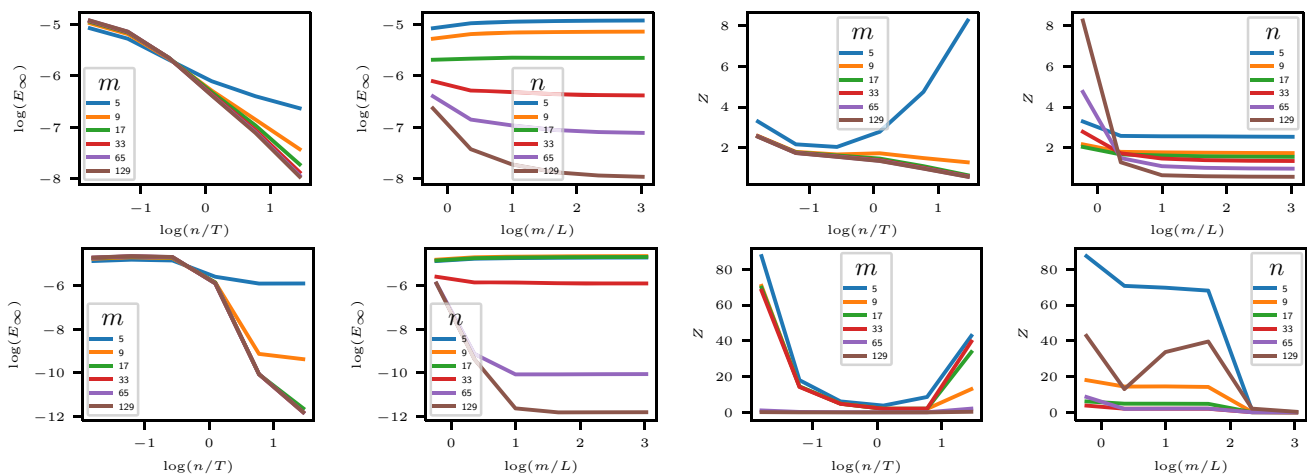
Fig. 2 Homogeneous Burger's equation, default prior (top row) and alternative prior (bottom row). For each pair (n, m) of temporal (n) and spatial (m) grid sizes considered, we plot: (left) the error E_∞ for fixed m and varying n; (centre left) the error E_∞ for fixed n and varying m; (centre right) the Z-score for fixed m and varying n; (right) the Z-score for fixed n and varying m

lihood and length-scale parameters fixed at values $\rho_1 = 1$, $\rho_2 = 2$ (not optimised; based on a simple *post-hoc* visual check).

4.2.2 Choice of linearisation

The differential operator here contains the nonlinear component $Qu = (\partial_x u)^2 + u \partial_x^2 u$ that must be linearised. The first term $(\partial_x u)^2$ appeared also in Burger's equation, and we linearise this term in an identical way to that used in Sect. 4.1. The second term $u \partial_x^2 u$ can be linearised in at least two distinct ways, fixing either u or $\partial_x^2 u$ to suitable constant values adaptively based on quantities that have been pre-computed. Thus we consider the two linearisations

$$Q_i^{(1)} u(t_i, x) := \mu^{i-1}(t_i, x) \frac{\partial u}{\partial x}(t_i, x) + \mu^{i-1}(t_i, x) \frac{\partial^2 u}{\partial x^2}(t_i, x)$$

$$Q_i^{(2)} u(t_i, x) := \mu^{i-1}(t_i, x) \frac{\partial u}{\partial x}(t_i, x) + u(t_i, x) \frac{\partial^2 \mu^{i-1}}{\partial x^2}(t_i, x)$$

where we recall that $\mu^{i-1}(t_i, x)$ is the predictive mean arising from the Gaussian process approximation U^{i-1}. Through simulation we aim to discover which (if either) linearisation is more appropriate for use in our PNM.

4.2.3 Conservation of mass

In addition to admitting multiple linearisations, we consider the porous medium equation because when $k > 1$ it exhibits a *conservation law*, which is typical of many nonlinear PDEs that are physically-motivated. Specifically, integrating Eq. (4.5) with respect to x gives

$$\frac{d}{dt} \int_{-L}^{L} u(t, x) \, dx - \left. \frac{\partial(u^k)}{\partial x} \right|_{-L}^{L} = 0$$

and, from the fact that $u = 0$ for all $x^2 \geq 12t^{2/3}$, it follows that $\partial_x(u^k) = ku^{k-1}\partial_x u = 0$ for all $x^2 \geq 12t^{2/3}$ and thus $\int_{-L}^{L} u(t, x) dx$ is t-invariant. A desirable property of a numerial method is that it respects conservations law of this kind; as exemplified by the *finite volume* methods (LeVeque 2002) and *symplectic integrators* (Sanz-Serna 1992). Interestingly, it is quite straight-forward to enforce this conservation law in our PNM by adding additional linear constraints to the system in 2.1. Namely, we add the linear constraints

$$\int_{-L}^{L} u(t_i, x) dx = \int_{-L}^{L} u(t_0, x) dx = 4(3^{\frac{1}{2}} - 3^{-\frac{1}{2}})$$

at each point $i \in \{1, \ldots, n-1\}$ on the temporal grid. The performance of our PNM both with and without conservation of mass will be considered.

4.2.4 Results

Empirical results based on the linearisation $Q^{(1)}$ (without conservation of mass) are contained in the top row of Fig. 3. In this case (and in contrast to our results for Burger's equation in Sect. 4.1), the error E_∞ is seen to be gated by the smaller of the finite length n of the temporal grid and the length m of the spatial grid. The Z-score values appear to be of order 1 as (n, m) are simultaneously increased, but are slightly higher than for Burger's equation, which may reflect the fact that the solution to the porous medium equation is only piecewise smooth. For increasing n with m fixed the PNM appears to become over-confident, while for increasing m with n fixed
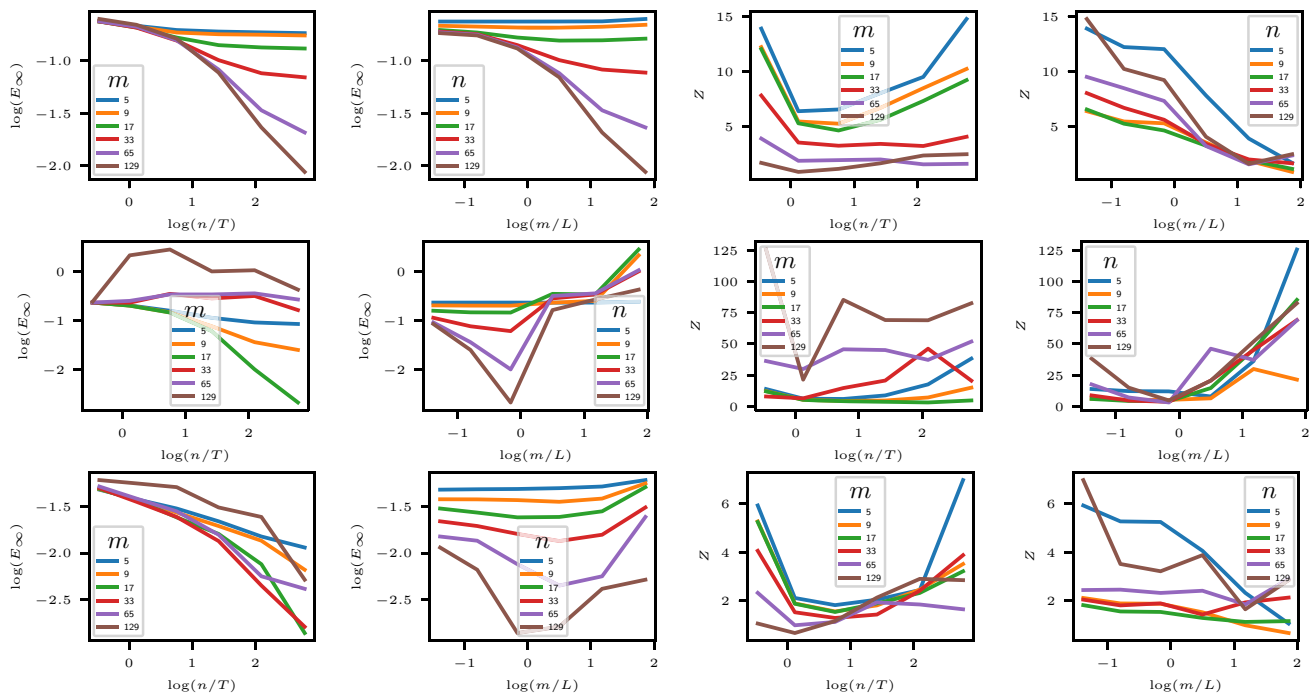
Fig. 3 Porous medium equation, with linearisation $Q^{(1)}$ (top row), with linearisation $Q^{(2)}$ (middle row), and with linearisation $Q^{(1)}$ and mass conserved (bottom row). For each pair (n, m) of temporal (n) and spatial (m) grid sizes considered, we plot: (left) the error E_∞ for fixed m and varying n; (centre left) the error E_∞ for fixed n and varying m; (centre right) the Z-score for fixed m and varying n; (right) the Z-score for fixed n and varying m

the PNM appears to become under-confident; a conservative choice would therefore be to take $m \geq n$. Interestingly, this behaviour of the Z-scores is similar to that observed for the rational quadratic covariance model in Sect. 4.1, and may reflect the fact that in both cases the solution u is outside the support of the covariance model. Next we compared the performance of the linearisation $Q^{(1)}$ with the linearisation $Q^{(2)}$. The error E_∞ associated to $Q^{(2)}$ (not shown) was larger than the error of $Q^{(1)}$, and the Z-scores for $Q^{(2)}$ are displayed in the middle row of Fig. 3. Our objective is to quantify numerical uncertainty, so it is essential that output from the PNM is calibrated. Unfortunately, it can be seen that the Z-scores associated with $Q^{(2)}$ are unsatisfactory; for large m the scores are two orders of magnitude larger than 1, indicating that the PNM is over-confident. The failure of $Q^{(2)}$ to provide calibrated output is likely due to the fact that approximation of the second-order derivative term $\partial_x^2 u$ is more challenging compared to approximation of the solution u, since ∂_x^2 is less regular than u and since our initial and boundary data relate directly to u itself.

Finally we considered inclusion of the conservation law into the PNM. For this purpose we used the best-performing linearisation $Q^{(1)}$. The errors E_∞ and Z-scores are shown in the bottom row of Fig. 3 and can be compared to the equivalent results without the conservation law applied, in the top row of Fig. 3. It can be seen that the error E_∞ is lower when the conservation law is applied, and moreover the Z-

scores are slightly reduced, remaining order 1. These results agree with the intuition that incorporating additional physical constraints, when they are known, can have a positive impact on the performance of our PNM.

4.3 Forced Burger's equation

Our final experiment concerns a nonlinear PDE whose right hand side f is considered to be a black box, associated with a substantial computational cost. To avoid confounding due to the choice of differential operator, we consider again the differential operator from Burger's equation

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} - \alpha \frac{\partial^2 u}{\partial x^2} = f(t, x), \ t \in [0, T], \ x \in [0, L],$$
(4.6)

for which the behaviour of our PNM was studied in Sect. 4.1 (in the case $f = 0$). The initial and boundary conditions are

$$u(0, x) = 0, \qquad\qquad x \in [0, L]$$
$$u(t, 0) = u(t, L) = 0, \qquad t \in [0, T]$$

and we set $\alpha = 1$, $T = 30$, $L = 1$. The aims of this experiment are two-fold: Our first aim is to evaluate the performance of our PNM when the function f is non-trivial (e.g. involving oscillatory behaviour), to understand whether the

output from our PNM remains calibrated or not. Recall that our experiments are *synthetic*, meaning that the black box f is in actual fact an analytic expression, in this case

$$f(t, x) := 10 \sin(6\pi x) \cos(3\pi t) + 2 |\sin(3\pi x) \cos(6\pi t)|,$$

enabling a thorough assessment to be performed. This forcing term is deliberately chosen to have some non-smoothness (from the absolute value function) and oscillatory behaviour, as might be encountered in output from a complex computer model. Our second aim is to compare the accuracy of our PNM against a classical numerical method whose computational budget (as quantified by the number of times f is evaluated) is identical to our PNM. The solution to Eq. (4.6) does not admit a closed form, so for our ground truth we used a numerical solution computed using the MATLAB function pdepe, which implements Skeel and Berzins (1990) based on a uniform spatial grid of size 512 and an adaptively selected temporal grid. Our PNM was implemented with the same linearisation used in Sect. 4.1.

4.3.1 Prior

Again we consider the default prior advocated in Sects. 3 and 4.1, with amplitude σ estimated using maximum likelihood and length-scale parameters fixed at values $\rho_1 = 0.5$, $\rho_2 = 0.5$ (not optimised; based on a simple *post-hoc* visual check).

4.3.2 Crank–Nicolson benchmark

In this scenario, where the black box function f is associated with a high computational cost, non-adaptive numerical methods are preferred, to control the total computational cost. From a classical perspective, the finite difference methods are natural candidates for the numerical solution of Eq. (4.6). Finite difference methods are classified into explicit and implicit schemes. Explicit schemes are much easier to solve but typically require certain conditions to be met for numerical stability (Thomas 1998, Table 5.3.1). For example, for the 2D heat equation, it is required that $\Delta t/(\alpha \Delta x^2) + \Delta t/(\alpha \Delta y^2) \leq 1/2$, where α is the diffusivity constant, Δt the time resolution, and Δx, Δy the spatial resolutions (Thomas 1998, page 158). Such conditions, which require the spatial resolution to be much finer than the time resolution, may be difficult to establish when f is a black box or when manual selection of the solution grid is not possible. Implicit methods in general are more difficult to solve, but stability is often guaranteed. For example, for the same 2D heat equation problem, the *Crank–Nicolson* scheme (a second-order, implicit method) is unconditionally stable (Thomas 1998, page 159). For these reasons, we considered the *Crank–Nicolson* finite difference method (Crank and Nicolson 1947) as a classical numerical method that is well-suited to the

task at hand. In the implementation of Crank-Nicholson on the inhomogeneous Burger's equation, the nonlinear term is approximated via a *lag nonlinear term* (Thomas, 1998, page140). The same regular temporal grid **t** and regular spatial grid **x** were employed in both Crank–Nicolson and our PNM, so that the computational costs for both methods (as quantified in terms of the number of evaluations of f) are identical.

4.3.3 Results

The error E_∞ and Z-scores for our PNM are displayed in Fig. 4. The error E_∞ is seen to be gated by the size m of the spatial grid and decreases as (n, m) are simultaneously increased. The Z-score values appear to be of order 1 as (n, m) are simultaneously increased, but for increasing n with m fixed the PNM appears to become over-confident; a conservative choice would be to take $m \geq n$, which is also what we concluded from the porous medium equation. These results suggest the output from our PNM is reasonably well-calibrated. Finally, we considered the accuracy of our PNM compared to the Crank–Nicolson benchmark. The error E_∞ for Crank–Nicolson is jointly displayed in Fig. 4, below the error plots for our PNM, and interestingly, it is generally larger than the error obtained with our PNM. This provides reassurance that our PNM is as accurate as could reasonably be expected.

5 Conclusion

This paper addressed an important and under-studied problem in numerical analysis; the numerical solution of a PDE under severe restrictions on evaluation of the initial, boundary and/or forcing terms f, g and h in Eq. (2.1). Such restrictions occur when f, g and/or h are associated with a computational cost, such as being output from a computationally intensive computer model (Fulton 2010; Hurrell et al. 2013) or arising as the solution to an auxiliary PDE (MacNamara and Strang 2016; Cockayne and Duncan 2020). In many such cases it is not possible to obtain an accurate approximation of the solution of the PDE, and at best one can hope to describe trajectories that are compatible with the limited information available on the PDE. To provide a principled resolution, in this paper we cast the numerical solution of a nonlinear PDE as an inference problem within the Bayesian framework and proposed a probabilistic numerical method (PNM) to infer the unknown solution of the PDE. This approach enables formal quantification of *numerical uncertainty*, in such settings where the solution of the PDE cannot be easily approximated.

Our contribution extends an active line of research into the development of PNM for a range of challenging numerical tasks (see the survey in Hennig et al. 2015,). A common
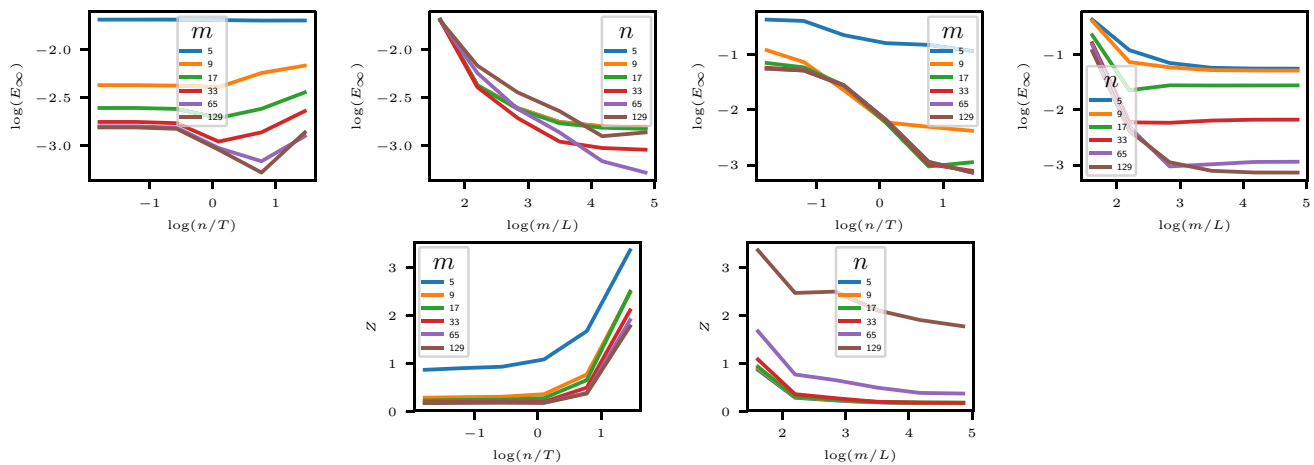
Fig. 4 Forced Burger's equation: For each pair (n, m) of temporal (n) and spatial (m) grid sizes considered, we plot: (top left) the error E_∞ for fixed m and varying n for our PNM; (top centre left) the error E_∞ for fixed n and varying m for our PNM; (top centre right) the error E_∞ for fixed m and varying n, Crank–Nicolson method; (top right right) the error E_∞ for fixed n and varying m, Crank–Nicolson method; (bottom left) the Z-score for fixed m and varying n; (bottom right) the Z-score for fixed n and varying m

feature of these tasks is that their difficulty justifies the use of sophisticated statistical machinery, such as Gaussian processes, that themselves may be associated with a computational cost. The PNM developed in this paper has a complexity $O(nm^3)$ to approximate the final state of the PDE, or $O(n^3 m^3)$ to approximate the full solution trajectory of the PDE. This renders our PNM computationally intensive – potentially orders of magnitude slower than a classical numerical method – but such an increase in cost can be justified when the demands of evaluating f, g and h exceed those of running the PNM (for example, when evaluation of f requires simulation from a climate model Fulton 2010; Hurrell et al. 2013,).

Further work will be required to establish our approach as a general-purpose numerical tool for nonlinear PDEs: First, the non-unique partitioning of the differential operator D into linear and nonlinear components, P and Q, together with the non-unique linearisation of Q, necessitates some expert input. This is analogous to the selection of a suitable numerical method in the classical setting, but the classical literature has benefited from decades of research and extensive practical guidance is now available in that context. Here we took a first step to automation by rigorously establishing sample path properties of a Matérn tensor product covariance model Σ, along with presenting a closed-form maximum likelihood estimator for the amplitude of Σ. The user is left to provide suitable length-scale parameter(s), which is roughly analogous to requiring the user to specify a mesh density in a finite element method (an accepted reality in that context). Second, an extensive empirical assessment will be required to systematically assess the performance of the method; our focus in the present paper was methodology and theory, pro-

viding only an experimental proof-of-concept. In particular, it will be important to assess diagnostics for *failure* of the method; it seems plausible that statistically-motivated diagnostics, such as held-out predictive likelihood, could be used to indicate the quality of the output from the PNM. Finally, we acknowledge that the problem we considered in Eq. (2.1) represents only one class of nonlinear PDEs and further work will be required to develop PNM for other classes of PDEs, such as boundary value problems and PDEs defined on more general domains.

Proof of Proposition 3.1

Proof (Proof of Proposition 3.1) The assumed regularity of μ, being an element of $C^p(I)$, implies that X and $X - \mu$ have identical differentiability properties up to order p, and we therefore assume $\mu = 0$ for simplicity in the remainder.

The mean square differentiability of the Matérn covariance function has been well-documented. In particular, because of the stationarity of the Matérn covariance function, X is order p mean-square differentiable if and only if $K_\nu^{(p,p)}(0,0) = (-1)^p K_\nu^{(2p)}(0)$ exists and is finite, and the Matérn covariance function with parameter ν is $2p$ times differentiable if and only if $\nu > p$; see Section 2 of Stein (1999). This establishes existence of the mean-square derivative $\partial_{\mathrm{MS}}^{(p)} X$.

It remains to prove that $\partial_{\mathrm{MS}}^{(p)} X$ is mean-square continuous. From the discussion in Eq. (3.1), $\partial_{\mathrm{MS}}^{(p)} X$ is mean-square continuous if and only if its autocovariance function, $K_\nu^{(2p)}$, is continuous at 0. Let $h = z - z'$ and $K_\nu(h) = f(h)g(h)$ where

$$f(h) = \sigma^2 \exp\left(-\frac{|h|}{\rho}\right) \frac{p!}{(2p)!},$$

$$g(h) = \sum_{k=0}^p \frac{(2p-k)!}{(p-k)!k!} \left(\frac{2}{\rho}\right)^k |h|^k$$

so that, by Leibniz's generalised product rule, for $m \in \mathbb{N}_0$,

$$K_\nu^{(m)}(h) = \sum_{n=0}^m \binom{m}{n} f^{(m-n)}(h) g^{(n)}(h).$$

One can verify that, for $n \in \{0, 1, \ldots, m\}$,

$$f^{(m-n)}(h) = \begin{cases} \frac{(-1)^{m-n}}{\rho^{m-n}} \sigma^2 \exp(-\frac{h}{\rho}) \frac{p!}{(2p)!} & h > 0 \\ \frac{1}{\rho^{2p-n}} \sigma^2 \exp(-\frac{-h}{\rho}) \frac{p!}{(2p)!} & h < 0 \end{cases}$$

$$g^{(n)}(h) = \begin{cases} \sum_{k=0}^{p-n} \frac{(2p-k)!}{(p-k)!k!} (\frac{2}{\rho})^{n+k} h^k & h > 0, n \le p \\ (-1)^n \sum_{k=0}^{p-n} \frac{(2p-k)!}{(p-k)!k!} (\frac{2}{\rho})^{n+k} (-h)^k & h < 0, n \le p \\ 0 & h \ne 0, n > p \end{cases}$$

from which it follows that

$$\lim_{h \downarrow 0} K_\nu^{(2p)}(h) = \frac{p!}{(2p)!} \sigma^2 \sum_{n=0}^p \binom{2p}{n} \frac{(2p-n)!}{(p-n)!} \frac{(-1)^{2p-n}}{\rho^{2p-n}} \left(\frac{2}{\rho}\right)^n$$

$$= \frac{\sigma^2}{\rho^{2p}} \sum_{n=0}^p \binom{p}{n} (-1)^{2p-n} 2^n$$

$$= (-1)^p \frac{\sigma^2}{\rho^{2p}} (2-1)^p$$

$$= (-1)^p \frac{\sigma^2}{\rho^{2p}}$$

and an analogous calculation shows

$$\lim_{h \uparrow 0} K_\nu^{(2p)}(h) = (-1)^p \frac{\sigma^2}{\rho^{2p}}.$$

Finally, we must check that the value $K_\nu^{(2p)}(0)$ agrees with the two limits just derived. $K_\nu^{(2p-1)}(h)$ is continuously differentiable, so $K_\nu^{(2p-1)}(0) = 0$ because it is an odd function (as it is an odd derivative of an even function K_ν). Thus we have that

$$K_\nu^{(2p)}(0) = \lim_{h \to 0} \frac{K_\nu^{(2p-1)}(h) - K_\nu^{(2p-1)}(0)}{h} = \lim_{h \to 0} \frac{K_\nu^{(2p-1)}(h)}{h}$$

$$= \lim_{h \to 0} \frac{p!}{(2p)!} \sigma^2 \exp\left(-\frac{|h|}{\rho}\right) \sum_{n=0}^{p-1} \binom{2p-1}{n}$$

$$\frac{(2p-n-1)!}{(p-n-1)!} \frac{(-1)^{2p-1-n}}{\rho^{2p-1-n}} \left(\frac{2}{\rho}\right)^{n+1} + O(h)$$

$$= (-1)^p \sigma^2 \sum_{n=0}^{p-1} \binom{p-1}{n} \frac{(-1)^{p-1-n} 2^n}{\rho^{2p}}$$

$$= (-1)^p \frac{\sigma}{\rho^{2p}} (2-1)^{p-1} = (-1)^p \frac{\sigma}{\rho^{2p}}$$

as required. $\qquad\square$

Proof of Proposition 3.2

As in the proof of Proposition 3.1, the assumed regularity of μ, being an element of $C^p(I)$, implies that X and $X - \mu$ have identical differentiability properties up to pth order, and we may therefore assume $\mu = 0$.

Our main tool is the Kolmogorov continuity theorem (see for example Kunita (1997), Section 1.4):

Theorem 6.1 *(Kolmogorov's Continuity Theorem) Let $I \subseteq \mathbb{R}^d$ be an open set, and let $\mathbf{z}$ be a dense subset of I. Let $X : I \times \Omega \to \mathbb{R}$ be a random field. If there exists constants $\alpha, \beta, C > 0$ such that*

$$\mathbb{E}\left[|X(z, \omega) - X(z', \omega)|^\alpha\right] \le C \|z - z'\|^{d+\beta}$$

for all $z, z' \in \mathbf{z}$, then there exists a modification of X that is sample continuous.

Lemma 6.1 *Let $I \subseteq \mathbb{R}^d$ be an open set and suppose that a positive definite function $\Sigma : I \times I \to \mathbb{R}$ satisfies*

$$\Sigma(z, z) + \Sigma(z', z') - 2\Sigma(z, z') \le C \|z - z'\|^\gamma$$

for some $\gamma, C \in (0, \infty)$ and all $z, z' \in I$. Let $X \sim \mathcal{GP}(0, \Sigma)$. Then there exists a modification of X that is sample continuous.

Proof (Proof of lemma 6.1) Notice that

$$\Sigma(z, z) + \Sigma(z', z') - 2\Sigma(z, z')$$
$$= \mathbb{E}[(X(z, \omega) - X(z', \omega))^2],$$

so if $\gamma > d$ then the required result follows from Kolmogorov's continuity theorem (6.1, with $\alpha = 2$). If not, then we can consider higher order moments via Isserlis's theorem: For $n \in \mathbb{N}$,

$$\mathbb{E}[(X(z, \omega) - X(z', \omega))^{2n}]$$
$$= \frac{(2n)!}{2^n n!} (\mathbb{E}[(X(z, \omega) - X(z', \omega))^2])^n$$

and thus, with any $n > d/\gamma$, we have

$$\mathbb{E}[(X(z, \omega) - X(z', \omega))^{2n}]$$
$$\leq \frac{(2n)!}{2^n n!} C^n \|z - z'\|^{\gamma n} \leq \tilde{C} \|z - z'\|^{d+\beta}$$

with $\tilde{C} = \frac{(2n)!}{2^n n!} C^n$ and $\beta = \gamma n - d$. The result then follows from Kolmogorov's continuity theorem (6.1, with $\alpha = 2n$). □

Proof (Proof of Proposition 3.2) Our aim is to show that $\partial_{\text{MS}}^{(i)} X$ satisfies the preconditions of 6.1. This process has covariance function $K_\nu^{(i,i)}(z, z') = (-1)^i K_\nu^{(2i)}(z - z')$. From stationarity we have, with $h = z - z'$,

$$K_\nu^{(2i)}(z, z) + K_\nu^{(2i)}(z', z') - 2K_\nu^{(2i)}(z, z')$$
$$= 2K_\nu^{(2i)}(0) - 2K_\nu^{(2i)}(h).$$

From similar calculations to those performed in the proof of 3.1, we have that for all $0 \leq i \leq p$,

$$K_\nu^{(2i)}(h) = \frac{2\sigma^2 p!}{(2p)!} \exp\left(-\frac{|h|}{\rho}\right) \sum_{n=0}^{\min(2i,p)}$$
$$\sum_{k=0}^{p-n} \binom{2i}{n} \frac{(2p-n-k)!}{(p-n-k)!k!} \frac{(-1)^{2i-n+1}}{\rho^{2i-n}} \left(\frac{2}{\rho}\right)^{n+k} |h|^k$$
$$= \exp\left(-\frac{|h|}{\rho}\right) \sum_{k=0}^{p} a_k |h|^k$$

for some real coefficients a_k. Therefore:

$$|2K_\nu^{(2i)}(0) - 2K_\nu^{(2i)}(h)|$$
$$= 2\left| a_0 - a_0 \exp\left(-\frac{|h|}{\rho}\right) - \exp\left(-\frac{|h|}{\rho}\right) \sum_{k=1}^{p} a_k |h|^k \right|$$

$$\leq 2\left| a_0 - a_0 \exp\left(-\frac{|h|}{\rho}\right) \right| + 2\left| \exp\left(-\frac{|h|}{\rho}\right) \sum_{k=1}^{p} a_k |h|^k \right|$$

$$(6.1)$$

This final term can be upper bounded by an expression of the form $C|h|^\gamma$ for sufficiently large $C > 0$ and $\gamma = 1$. Indeed, as $h \to 0$ the behaviour of Eq. (6.1) is $O(|h|)$. As $|h| \to \infty$ the exponential term dominates and Eq. (6.1) decays to a_0. In the region $0 < |h| < \infty$, Eq. (6.1) is smooth. Thus we can use 6.1 to conclude that $\partial_{\text{MS}}^{(i)} X$ has a modification that is sample continuous. □

Proof of Corollary 3.1

For $p = 1$ the result is immediate from 3.1, so in what follows we concentrate on $p > 1$.

The main technical challenge of this proof is to deal with modifications, which arise with each application of 3.1. Recall that two stochastic processes X and $\tilde{X}$ are said to be *indistinguishable* if $\mathbb{P}(X(z, \omega) = \tilde{X}(z, \omega) \; \forall z \in I) = 1$. If X and $\tilde{X}$ are modifications of each other and each is sample continuous, then X and $\tilde{X}$ are indistinguishable (Jeanblanc et al. 2009, Section 1.1).

Proof (Proof of Corollary 3.1) We first present a proof for $d = 1$, to improve transparency of the argument, then we present the argument for the general case $d \geq 1$. Note that, since we are considering Gaussian processes, the requirement for a second moment in 3.1 is automatically satisfied.

For each $0 \leq i < p$, it is assumed that $\partial_{\text{MS}}^i X$ has mean square derivative $\partial_{\text{MS}}^{i+1} X$ that is mean square continuous and sample continuous. 3.1 therefore implies that each $\partial_{\text{MS}}^i X$ has a modification, denoted ψ_i, that is sample continuously differentiable, and satisfying $\partial \psi_i = \partial_{\text{MS}}^{i+1} X$ almost surely. Since ψ_i and $\partial_{\text{MS}}^i X$ are sample continuous they are indistinguishable; i.e. almost surely $\psi_i = \partial_{\text{MS}}^i X$. It follows that, for each $0 \leq i < p$, we have almost surely that $\partial^i \psi_0 = \partial^{i-1}(\partial \psi_0) = \partial^{i-1} \psi_1 = \cdots = \psi_i$, while for $i = p$ we have that $\partial^p \psi_0 = \partial \psi_{p-1} = \partial_{\text{MS}}^p X$.

The case $d \geq 1$ is analogous with more notation is involved; though, since we assumed $\Sigma \in C^{(p,p)}(I \times I)$ we may employ the shorthand notation $\partial_{\text{MS}}^\beta X$ for all $|\beta| \leq p$ (since the order of derivatives can be freely interchanged). For each $0 \leq i < p$ and $|\beta| = i$, it is assumed that $\partial_{\text{MS}}^\beta X$ has mean square partial derivatives $\partial_{\text{MS}}^{\beta+\gamma} X$, $|\gamma| = 1$, that are mean square continuous and sample continuous. 3.1 therefore implies that each $\partial_{\text{MS}}^\beta X$ has a modification, denoted ψ_β, that is sample continuously differentiable, and satisfying $\partial^\gamma \psi_\beta = \partial_{\text{MS}}^{\beta+\gamma} X$ almost surely for all $|\gamma| = 1$. Since ψ_β and $\partial_{\text{MS}}^\beta X$ are sample continuous they are indistinguishable; i.e. almost surely $\psi_\beta = \partial_{\text{MS}}^\beta X$. It follows that, for each $0 \leq i < p$ and $|\beta| = i$, we have almost surely that $\partial^\beta \psi_0 = \psi_\beta$, while for and $|\beta| = p$ we have that $\partial^\beta \psi_0 = \partial_{\text{MS}}^\beta X$. □

Proof of Theorem

The following technical result will be exploited:

Theorem 6.2 *(Integrated Gaussian Process; Corollary 1 and 2 of Lasinger (1993))*

Let $I = (0, T)$ and let $X : I \times \Omega \to \mathbb{R}$ be a stationary Gaussian process with mean function $\mu = 0$ and autocovariance function Σ. Let $q \in \mathbb{N}$ and recursively define $X^{(-i)}$ for $i = 1, \ldots, q$ as follows:

$$X^{(-i)}(t) = X^{(-i)}(0) + \int_0^t X^{(-i+1)}(t') \, dt'$$

Let A_i, $i = 1, \ldots, q$ be constants and recursively define $\Sigma^{(-i,-i)}$ for $i = 1, \ldots, q$ as follows:

$$\Sigma^{(-i,-i)}(t) = A_i^2 - \int_0^t \int_0^s \Sigma^{(-i+1,-i+1)}(s') \, ds' ds.$$

Let $\mathcal{H}(\Sigma)$ denote the reproducing kernel Hilbert space associated with the covariance function Σ, and $\mathcal{H}(\Sigma^{(-i,-i)})$ the reproducing kernel Hilbert space associated with the covariance function $\Sigma^{(-i,-i)}$. If

$$\partial^{q+i-1} \Sigma^{(-q,-q)}(\cdot) \in \mathcal{H}(\Sigma) \tag{6.2}$$

and

$$A_i^2 > \|\partial \Sigma^{(-i,-i)}(\cdot)\|_{\mathcal{H}(\Sigma^{(-(i-1),-(i-1))})}^2 \tag{6.3}$$

for each $1 \leq i \leq q$, then there exists a stationary Gaussian process $X^{(-q)} : I \times \Omega \to \mathbb{R}$ with covariance function $\Sigma^{(-q,-q)}(t, s) := \Sigma^{(-q,-q)}(t - s)$ such that $\partial^q \bar{\partial}^q \Sigma^{(-q,-q)}(t, s) = \Sigma(t - s)$. Moreover, Eq. (6.2) implies $\partial \Sigma^{(-i,-i)}(\cdot) \in \mathcal{H}(\Sigma^{(-(i-1),-(i-1))})$, so that A_i satisfying Eq. (6.3) exist.

From 6.2 we obtain the following intermediate result:

Proposition 6.1 *Let $I = (0, T)$ and let $\mu \in C^p(I)$ be bounded in I. Let $X \sim \mathcal{GP}(\mu, K_\nu)$, with K_ν as in Eq. (3.2). Let $q \in \{0, \ldots, p\}$. Then there exists a stationary Gaussian process $X^{(-q)} : I \times \Omega \to \mathbb{R}$ whose mean function $\mu^{(-q)}$ satisfies $\partial^q \mu^{(-q)}(t) = \mu(t)$ and whose covariance function $K_\nu^{(-q,-q)}(\cdot)$ satisfies $\partial^q \bar{\partial}^q K_\nu^{(-q,-q)}(t, s) = K_\nu(t, s)$. Furthermore, $X^{(-q)}$ is sample continuous.*

Proof Since $\mu \in C^p(I)$ we may define $\mu^{(-q)}$ to be the q times integrated μ for each $q \in \{0, \ldots, p\}$. Note that $\mu^{(-q)}$ is well-defined since μ is assumed to be bounded.

Our aim is to use 6.2, so we must check that the preconditions are satisfied. Eq. (6.2) is satisfied because H_K for the Matérn ν kernel is norm equivalent to the Sobolev Space W_2^α where $\alpha = \nu + 1/2$ () and integrals of W_2^α functions are W_2^α functions. Equation (6.3) is automatically satisfied since we may select suitably large A_i^2. Thus 6.2 is satisfied and there exists a stationary Gaussian process

$X^{(-q)} : I \times \Omega \to \mathbb{R}$ whose mean function $\mu^{(-q)}$ satisfies $\partial^q \mu^{(-q)}(t) = \mu(t)$ and whose covariance function $K_\nu^{(-q,-q)}(\cdot)$ satisfies $\partial^q \bar{\partial}^q K_\nu^{(-q,-q)}(t, s) = K_\nu(t, s)$.

It remains only to show that there is a modification of $X^{(-q)}$ is sample continuous. Since $K_\nu(t)$ is of the form $K_\nu(t) = \exp(-t/\rho) \sum_{k=0}^p a_k t^k$ for some real constants a_k, we have

$$K_\nu^{(-q,-q)}(t) = \exp\left(-\frac{t}{\rho}\right) \sum_{k=0}^p b_k t^k + \sum_{k=0}^{2q-1} c_k t^k$$

for some real constants b_k and c_k. Now we check the preconditions of 6.1:

$$K_\nu^{(-q,-q)}(t, t) + K_\nu^{(-q,-q)}(s, s) - 2K_\nu^{(-q,-q)}(t, s)$$
$$= 2K_\nu^{(-q,-q)}(0) - 2K_\nu^{(-q,-q)}(|t - s|)$$
$$= 2b_0 - 2b_0 \exp\left(-\frac{|t - s|}{\rho}\right)$$
$$\quad - 2 \exp\left(-\frac{|t - s|}{\rho}\right) \sum_{k=1}^p b_k |t - s|^k$$
$$\quad - 2 \sum_{k=1}^{2q-1} c_k |t - s|^k$$
$$\leq C|t - s|$$

on a bounded domain, for some constant C. Therefore from 6.1 there exists a modification of $X^{(-q)}$ that is sample continuous, as required. □

Finally we present the proof of 3.2. The main technical challenge is to smooth the process in such a way that it has an equal number of derivatives in each argument, to allow for a more straightforward application of earlier results.

Proof (Proof of Theorem 3.2) If the a_i, b_i are not equal to 0 and 1 respectively, we perform a linear re-scaling so that without loss of generality we can consider the hyper-rectangle $I = (0, 1)^d$ in the remainder.

Let $p = \max_i \beta_i$ and let $P \in \mathbb{N}_0^d$ denote the vector whose entries are all equal to p. Consider a stationary Gaussian process $Z : I \times \Omega \to \mathbb{R}$ with mean function $\mu^{(\beta - P)}(\cdot)$ and autocovariance function

$$\Sigma^{(\beta - P, \beta - P)}(t - s) = \prod_{i=1}^d K_{\nu_i}^{(\beta_i - p, \beta_i - p)}(t_i - s_i), \tag{6.4}$$

expressed using the notation of 6.2. This construction ensures that Z is order p mean-square differentiable, since $\mu^{(\beta - P)} \in C^p(I)$ and since Eq. (6.4) is a product of functions $K_{\nu_i}^{(\beta_i - p, \beta_i - p)}(\cdot) \in C^{(p,p)}(I \times I)$, as a consequence of Propositions 3.1 and 6.1.

Next we show that $\partial_{\mathrm{MS}}^{\alpha} Z$ has a sample continuous modification, for any $\alpha \in \mathbb{N}_0^d$, $|\alpha| \leq p$, using 6.1. To apply 6.1 it suffices to show that

$$\left| \partial^{\alpha} \bar{\partial}^{\alpha} \Sigma^{(\beta-P, \beta-P)}(0) - \partial^{\alpha} \bar{\partial}^{\alpha} \Sigma^{(\beta-P, \beta-P)}(t) \right| \leq C \|t\|^{\gamma} \tag{6.5}$$

for some $\gamma, C \in (0, \infty)$ and all $t \in I$. From 6.1, the first term on the left hand side satisfies $\partial^{\alpha} \bar{\partial}^{\alpha} \Sigma^{(\beta-P, \beta-P)}(t, s) = \Sigma^{(\beta+\alpha-P, \beta+\alpha-P)}(t, s)$ and thus Eq. (6.5) becomes

$$\left| \Sigma^{(\beta+\alpha-P, \beta+\alpha-P)}(0) - \Sigma^{(\beta+\alpha-P, \beta+\alpha-P)}(t) \right| \leq C \|t\|^{\gamma}. \tag{6.6}$$

Since $\beta_i + \alpha_i - P_i \leq \beta_i$, and from the product form of $\Sigma^{(\beta+\alpha-P, \beta+\alpha-P)}$ and the univariate calculations performed in the proofs of 3.2 and 6.1, Eq. (6.6) is verified. Thus, by 6.1, $\partial_{\mathrm{MS}}^{\alpha} Z$ has a sample continuous modification, for any $\alpha \in \mathbb{N}_0^d$, $|\alpha| \leq p$.

Finally, we apply 3.1 to deduce that there exists a modification $\tilde{Z}$ of Z with $\mathbb{P}(\tilde{Z} \in C^p(I))$. It follows that $\tilde{X} = \partial^{P-\beta} \tilde{Z}$ is a modification of X that satisfies $\mathbb{P}(\tilde{X} \in C^{\beta}(I))$. $\qquad\square$

References

Barenblatt, G.I.: On some unsteady motions of a liquid and gas in a porous medium. Akad. Nauk SSSR Prikl Mat. Meh. **16**, 67–78 (1952)

Bosch, N., Hennig, P., Tronarp, F.: Calibrated adaptive probabilistic ODE solvers. In: Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, PMLR, 130, pp 3466–3474 (2021)

Chen, J., Chen, Z., Zhang, C., Wu, C.F.J.: APIK: Active physics-informed kriging model with partial differential equations. http://arxiv.org/abs/201211798 (2020)

Chen, Y., Hosseini, B., Owhadi, H., Stuart, A.M.: Solving and learning nonlinear PDEs with Gaussian processes. http://arxiv.org/abs/210312959 (2021)

Chkrebtii, O.A., Campbell, D.A.: Adaptive step-size selection for state-space probabilistic differential equation solvers. Stat. Comput. **29**(6), 1285–1295 (2019). https://doi.org/10.1007/s11222-019-09899-5

Chkrebtii, O.A., Campbell, D.A., Calderhead, B., Girolami, M.A.: Bayesian solution uncertainty quantification for differential equations. Bayesian Anal. **11**(4), 1239–1267 (2016). https://doi.org/10.1214/16-BA1017

Cockayne, J., Duncan, A.B.: Probabilistic gradients for fast calibration of differential equation models. http://arxiv.org/abs/200904239 (2020)

Cockayne, J., Graham, M.M., Oates, C.J., Sullivan, T.J.: Testing whether a learning procedure is calibrated. http://arxiv.org/abs/201212670 (2021)

Cockayne, J., Oates, C., Sullivan, T., Girolami, M.: Probabilistic meshless methods for partial differential equations and Bayesian inverse problems. http://arxiv.org/abs/160507811 (2016)

Cockayne, J., Oates, C.J., Sullivan, T.J., Girolami, M.: Bayesian probabilistic numerical methods. SIAM Rev. **61**(4), 756–789 (2019). https://doi.org/10.1137/17M1139357

Crank, J., Nicolson, P.: A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. Proc. Cambridge Philos. Soc. **43**, 50–67 (1947)

de Roos, F., Gessner, A., Hennig, P.: High-dimensional Gaussian process inference with derivatives. http://arxiv.org/abs/210207542 (2021)

Diaconis, P.: Bayesian numerical analysis. Stat. Decis. Theor. Relat. Top., Vol IV **1**, 163–175 (1988)

Dudley, R.M.: The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. J. Funct. Anal. **1**, 290–330 (1967). https://doi.org/10.1016/0022-1236(67)90017-1

Evans, L.C.: Partial Differential Equations, Graduate Studies in Mathematics, vol. 19. American Mathematical Society, Providence, RI (1998)

Fasshauer, G.E.: Solving differential equations with radial basis functions: multilevel methods and smoothing. Adv. Comput. Math. **11**(2–3), 139–159 (1999). https://doi.org/10.1023/A:1018919824891

Fornberg, B., Flyer, N.: Solving PDEs with radial basis functions. Acta. Numer. **24**, 215–258 (2015). https://doi.org/10.1017/S0962492914000130

Fulton, E.A.: Approaches to end-to-end ecosystem models. J. Mar. Syst. **81**(1–2), 171–183 (2010)

Gneiting, T.: Compactly supported correlation functions. J. Multivar. Anal. **83**(2), 493–508 (2002). https://doi.org/10.1006/jmva.2001.2056

Grätsch, T., Bathe, K.J.: A posteriori error estimation techniques in practical finite element analysis. Comput. Struct. **83**(4–5), 235–265 (2005). https://doi.org/10.1016/j.compstruc.2004.08.011

Hennig, P., Osborne, M.A., Girolami, M.: Probabilistic numerics and uncertainty in computations. Proc. A **471**(2179), 20150142 (2015). https://doi.org/10.1098/rspa.2015.0142

Hurrell, J.W., Holland, M.M., Gent, P.R., Ghan, S., Kay, J.E., Kushner, P.J., Lamarque, J.F., Large, W.G., Lawrence, D., Lindsay, K., et al.: The community earth system model: a framework for collaborative research. Bull. Am. Meteorol. Soc. **94**(9), 1339–1360 (2013)

Jeanblanc, M., Yor, M., Chesney, M.: Mathematical methods for financial markets. Springer, Cham (2009)

Jidling, C., Wahlström, N., Wills, A., Schön, T.B.: Linearly constrained Gaussian processes. In: Proceedings of the 31st Conference on Neural Information Processing Systems (2017)

Kanagawa, M., Hennig, P., Sejdinovic, D., Sriperumbudur, B.K.: Gaussian processes and kernel methods: A review on connections and equivalences. http://arxiv.org/abs/180702582 (2018)

Karvonen, T.: Small sample spaces for gaussian processes. http://arxiv.org/abs/210303169 (2021)

Karvonen, T., Wynne, G., Tronarp, F., Oates, C., Särkkä, S.: Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions. SIAM/ASA J. Uncertain Quantif. **8**(3), 926–958 (2020). https://doi.org/10.1137/20M1315968

Kennedy, M.C., O'Hagan, A.: Bayesian calibration of computer models. J. R. Stat. Soc. Ser. B Stat. Methodol. **63**(3), 425–464 (2001). https://doi.org/10.1111/1467-9868.00294

Kersting, H., Hennig, P.: Active uncertainty calibration in Bayesian ODE solvers. In: Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016), pp 309–318 (2016)

Kersting, H., Sullivan, T.J., Hennig, P.: Convergence rates of Gaussian ODE filters. Stat. Comput. **30**(6), 1791–1816 (2020). https://doi.org/10.1007/s11222-020-09972-4

Krämer, N., Hennig, P.: Stable implementation of probabilistic ODE solvers. http://arxiv.org/abs/201210106 (2020)

Kunita, H.: Stochastic Flows and Stochastic Differential Equations, Cambridge Studies in Advanced Mathematics, vol 24. Cambridge University Press, Cambridge, reprint of the 1990 original (1997)

Larkin, F.M.: Gaussian measure in Hilbert space and applications in numerical analysis. Rocky Mountain J. Math. **2**(3), 379–421 (1972). https://doi.org/10.1216/RMJ-1972-2-3-379

Lasinger, R.: Integration of covariance kernels and stationarity. Stoch. Process Appl. **45**(2), 309–318 (1993). https://doi.org/10.1016/0304-4149(93)90077-H

LeVeque, R.J.: Finite Volume Methods for Hyperbolic Problems. Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, (2002). https://doi.org/10.1017/CBO9780511791253

MacNamara, S., Strang, G.: Operator splitting. splitting methods in communication. Imaging, science, and engineering, pp. 95–114. Springer, Cham (2016)

Marcus, M.B., Shepp, L.A.: Sample behavior of gaussian processes. In: Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory, University of California Press, Berkeley, Calif., pp 423–441 (1972)

Oates, C.J., Sullivan, T.J.: A modern retrospective on probabilistic numerics. Stat. Comput. **29**(6), 1335–1351 (2019). https://doi.org/10.1007/s11222-019-09902-z

Oates, C.J., Cockayne, J., Aykroyd, R.G., Girolami, M.: Bayesian probabilistic numerical methods in time-dependent state estimation for industrial hydrocyclone equipment. J. Am. Stat. Assoc. **114**(528), 1518–1531 (2019). https://doi.org/10.1080/01621459.2019.1574583

O'Hagan, A.: Some Bayesian numerical analysis Bayesian statistics, vol. 4. Oxford University Press, Cambridge (1992)

Owhadi, H.: Bayesian numerical homogenization. Multiscale Model. Simul. **13**(3), 812–828 (2015)

Petzold, L., Li, S., Cao, Y., Serban, R.: Sensitivity analysis of differential-algebraic equations and partial differential equations. Computers Chemical Eng. **30**(10–12), 1553–1559 (2006)

Potthoff, J.: Sample properties of random fields III: differentiability. Commun. Stoch. Anal. **4**(3), 335–353 (2010). https://doi.org/10.31390/cosa.4.3.03

Raissi, M., Perdikaris, P., Karniadakis, G.E.: Numerical Gaussian processes for time-dependent and nonlinear partial differential equations. SIAM J. Sci. Comput. **40**(1), A172–A198 (2018). https://doi.org/10.1137/17M1120762

Rasmussen, C.E., Williams, C.K.: Gaussian processes for machine learning. The MIT Press, Cambridge (2006)

Sanz-Serna, J.M.: Symplectic integrators for Hamiltonian problems: An overview. Acta Numer. **1**, 243–286 (1992)

Särkkä, S.: Linear operators and stochastic partial differential equations in Gaussian process regression. In: International Conference on Artificial Neural Networks, Springer, pp 151–158 (2011)

Schäfer, F., Sullivan, T.J., Owhadi, H.: Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity. Multiscale Model. Simul. **19**(2), 688–730 (2021). https://doi.org/10.1137/19M129526X

Scheuerer, M.: Regularity of the sample paths of a general second order random field. Stoch. Process. Appl. **120**(10), 1879–1897 (2010)

Schober, M., Duvenaud, D.K., Hennig, P.: Probabilistic ODE solvers with Runge–Kutta means. In: Proceedings of the 28th Annual Conference on Neural Information Processing Systems, pp 739–747 (2014)

Schober, M., Särkkä, S., Hennig, P.: A probabilistic model for the numerical solution of initial value problems. Stat. Comput. **29**(1), 99–122 (2019). https://doi.org/10.1007/s11222-017-9798-7

Schölkopf, B., Herbrich, R., Smola, A.J.: A generalized representer theorem. In: Proceedings of the 14th International Conference on Conference on Computational Learning Theory, Springer, pp 416–426 (2001)

Skeel, R.D., Berzins, M.: A method for the spatial discretization of parabolic equations in one space variable. SIAM J. Sci. Stat. Comput. **11**(1), 1–32 (1990). https://doi.org/10.1137/0911001

Skilling, J.: Bayesian solution of ordinary differential equations. maximum entropy and Bayesian methods. Springer, New York (1992)

Stein, M.L.: Interpolation of Spatial Data. Springer Series in Statistics, Springer-Verlag, New York, (1999). https://doi.org/10.1007/978-1-4612-1494-6

Strikwerda, J.C.: Finite Difference Schemes and Partial Differential Equations, 2nd edn. Society for Indistrial and Applied Mathematics (2004)

Teymur, O., Lie, H.C., Sullivan, T., Calderhead, B.: Implicit probabilistic integrators for ODEs. In: Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (2018)

Teymur, O., Zygalakis, K., Calderhead, B.: Probabilistic linear multistep methods. In: Proceedings of the 30th Annual Conference on Neural Information Processing Systems, pp 4321–4328 (2016)

Thomas, J.: Numerical partial differential equations: finite difference methods. Springer Science & Business Media, New York (1998)

Tronarp, F., Kersting, H., Särkkä, S., Hennig, P.: Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: a new perspective. Stat. Comput. **29**(6), 1297–1315 (2019). https://doi.org/10.1007/s11222-019-09900-1

Tronarp, F., Särkkä, S., Hennig, P.: Bayesian ODE solvers: the maximum a posteriori estimate. Stat. Comput. **31**(3), 23 (2021). https://doi.org/10.1007/s11222-021-09993-7

Wang, X., Berger, J.O.: Estimating shape constrained functions using Gaussian processes. SIAM/ASA J. Uncertain Quantif. **4**(1), 1–25 (2016)

Wang, J., Cockayne, J., Oates, C.J.: A role for symmetry in the Bayesian solution of differential equations. Bayesian Anal. **15**(4), 1057–1085 (2020). https://doi.org/10.1214/19-BA1183

Wheeler, M.W., Dunson, D.B., Pandalai, S.P., Baker, B.A., Herring, A.H.: Mechanistic hierarchical Gaussian processes. J. Am. Stat. Assoc. **109**(507), 894–904 (2014). https://doi.org/10.1080/01621459.2014.899234