

# PHYSICS-BASED COVARIANCE MODELS FOR GAUSSIAN PROCESSES WITH MULTIPLE OUTPUTS

*Emil M. Constantinescu\* & Mihai Anitescu*

*Mathematics and Computer Science Division, Argonne National Laboratory, 9600 S. Cass Avenue, Argonne, Illinois 60439, USA*

*Original Manuscript Submitted: 07/15/2011; Final Draft Received: 09/28/2011*

*Gaussian process analysis of processes with multiple outputs is limited by the fact that far fewer good classes of covariance functions exist compared with the scalar (single-output) case. To address this difficulty, we turn to covariance function models that take a form consistent in some sense with physical laws that govern the underlying simulated process. Models that incorporate such information are suitable when performing uncertainty quantification or inferences on multidimensional processes with partially known relationships among different variables, also known as cokriging. One example is in atmospheric dynamics where pressure and wind speed are driven by geostrophic assumptions ( $\text{wind} \propto \partial/\partial x \text{ pressure}$ ). In this study we develop both analytical and numerical auto-covariance and cross-covariance models that are consistent with physical constraints or can incorporate automatically sensible assumptions about the process that generated the data. We also determine high-order closures, which are required for nonlinear dependencies among the observables. We use these models to study Gaussian process regression for processes with multiple outputs and latent processes (i.e., processes that are not directly observed and predicted but inter-relate the output quantities). Our results demonstrate the effectiveness of the approach on both synthetic and real data sets.*

**KEY WORDS:** *Gaussian random field, spatial uncertainty, model calibration, spatial statistics*

## 1. INTRODUCTION

In this work, we explore Gaussian process (GP) regression [1–4] for models driven by physical principles in general and for regression and state estimations, in particular. Predictions and spatial interpolation (kriging) using GPs is a well-established technique [1, 5, 6]. Inferences on processes with multiple outputs, the topic of particular interest in this work, is known as cokriging [5, 7] or multikriging [8]. Here, we focus on spatial processes; nonetheless, GPs can also be used in a time-dependent context processes [9–13] or in a spatio-temporal context [14].

One of the important operational decisions in uncertainty quantification, and in particular in carrying out Gaussian process analysis, is the choice of covariance functions. For kriging for scalar (single-output) fields several well-understood practical and theoretical guidelines exist [3, 5]. However, with multiple outputs it is difficult to describe the process in order to correctly structure the outputs and ensure positive definiteness [3]. One approach is introduced by [8, 15] in which smoothing kernels are used to train how the outputs covary.

The difficulty of finding “good” covariance models for multiple outputs can have important practical consequences. An incorrect structure of the covariance matrix can significantly reduce the efficiency of the uncertainty quantification process, as well as the forecast efficiency in kriging inferences [16]. Therefore, we argue, the covariance model may play an even more profound role in cokriging [7, 17]. This argument applies when the covariance structure is inferred from data, as is typically the case. Such studies as the ones discussed in [7, 17] have been replicated by [18–20] and channeled toward constructing compact kernels, which are positive functions with compact support that avoid matrix storage issues, although no connection was made between the two sets of studies. For example, in [16] poor results were obtained when an isotropic model was used instead of a more appropri-

---

\*Correspond to Emil M. Constantinescu, E-mail: emconsta@mcs.anl.gov, URL: <http://www.mcs.anl.gov/~emconsta/>

ate anisotropic one. We expect the situation to be even more critical for systems with outputs that have “different” physical meanings. In this case the auto- and cross-covariance models [7] determine the efficiency of the kriging process.

The aim of this work is to obtain covariance functions for multivariate processes by using information about the physics of the process. The regime of interest here is the one where there exists sufficient information about the physics of the process to generate suitable covariance functions, but there is not enough information to use the mathematical equations directly (for example, boundary or initial conditions may be unknown, but observations of the process are available).

In statistics, physical intuition is often employed directly or indirectly in solving inference problems. A typical strategy to include *ab initio* knowledge about a real system governed by known (in part) physical laws is hierarchical Bayesian modeling [21, 22]. Specific examples include geophysical processes [7], atmospheric modeling [23–26], and environmental sciences [24, 27–30]. The physical component of the problem is typically oversimplified in order to allow tractable computations. Relevant studies for this work include [7, 18, 25, 26, 31–33], in which the authors include certain levels of particular physical properties in the covariance structure. Auto- and cross-covariance models induced from such processes present in hydrogeology or driven from stochastic differential equations are discussed in [7, 26]. In [34] correlation functions are inferred from specific linear partial differential equations (PDEs). Apanasovich et al. [35] propose cross-covariance functions for multivariate random fields obtained by a multidimensional extension of existing univariate models. Gneiting et al. [36] introduce multivariate Matérn cross-covariance functions that allow each process component to maintain different smoothness properties. Constructing covariance functions that preserve liquid incompressibility via the divergence operator is discussed in [37]. In this study, we extend and generalize previous results by providing a general framework for assembling consistent auto- and cross-covariance models that are asymptotically consistent with functional constraints that may depend both on the covariates and on observations. The setup comprises multiple observed outputs, with underlying processes that constrain the data. This setup also extends to processes that are not necessarily directly observed. We term the latter a hidden process model, which is a parallel to hidden (latent) Markov chain models.

In addition to deriving a systematic approach for describing the construction of covariance models governed by linear processes, we ask what happens if the process is not linear. We find that high-order closures are necessary to correctly specify the resulting covariance models. Moreover, the strategy that we introduce in this study provides a physically consistent approach to introduce nonstationarity in the structure of the covariance matrix.

In this study we focus on Bayesian linear regression with normally distributed forcings, where the response (vector) variables  $y_i$  are functions of covariates  $x$ , which in our case and without loss of generality are considered locations in space. In addition, we consider that there is a relationship among the  $m$  output fields, each of dimension  $n_i$ , that is driven by an underlying physical process:

$$\mathbf{0}_{nm} = f(y_1, y_2, \dots, y_m) + \psi, \quad (1)$$

where  $n = \sum_i n_i$ ,  $\mathbf{0}_{nm}$  is a vector of zeros,  $y_i \in \mathbb{R}^{n_i}$ ,  $f : \mathbb{R}^{nm} \rightarrow \mathbb{R}^{nm}$ , is a physical deterministic model that connects the different physical quantities  $y_i$ , and  $\psi$  is a stochastic forcing (random vector) that accounts for the difference between our knowledge of the physical process and the real one. One example is steady-state calculations for fluid dynamics problems, where  $y_i$  may represent mass, momentum, and energy; and  $\psi$  corresponds to the departure of the real model from the idealized one. Another example is given by imposing a divergence-free constraint on a multidimensional field, such as liquid incompressibility discussed in [37]. We call physical model (1) separable if we can write it as

$$\begin{aligned} y_1 &= f_1(y_1, y_2, \dots, y_m) + \psi_1 \\ y_2 &= f_2(y_1, y_2, \dots, y_m) + \psi_2 \\ &\dots = \dots \\ y_m &= f_m(y_1, y_2, \dots, y_m) + \psi_m \end{aligned}, \quad (2)$$

where  $[f_1^T, f_2^T, \dots, f_m^T]^T = f$  and  $[\psi_1^T, \psi_2^T, \dots, \psi_m^T]^T = \psi$ . This is not to be confused with separable covariance models. These situations occur, for instance, in implicit temporal discretization of PDEs, where the subscript

represents the time index. A separable model (2) is explicit if we can write it as  $y_i = f_i(y_{j_1}, y_{j_2}, \dots, y_{j_d})$ , where  $i = 1, \dots, m$ ,  $j_k \in \mathcal{J} \setminus \{i\}$  is an index set that does not contain  $i$ . In other words we can write (2) as

$$y_i = g_i(\{y_1, \dots, y_m\} \setminus \{y_i\}) + \psi, \quad (3)$$

where  $g$  is an explicit model of a known physical process that governs some  $y_i$  variables. Henceforth we will denote by  $g$  explicit functions as defined above. One example is the geostrophic wind approximation  $u = k \times (1/\rho c) \nabla p$  [38] where in (3) the wind speed  $u$  takes the role of  $y_1$ , and the pressure  $p$  of  $y_2$ . This example will be used later in this study. For a relevant use of gradients in kriging see [39].

Two types of models are considered in order to simplify the presentation of the theoretical discussion: an implicit separable one, which for exposition brevity will generally be defined on two random fields,

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = f(y_1, y_2) + \psi = \begin{bmatrix} f_1(y_1, y_2) + \psi_1 \\ f_2(y_1, y_2) + \psi_2 \end{bmatrix}, \quad (4)$$

and an explicit one defined by

$$y_1 = g(y_2) + \eta, \quad (5)$$

where we assume that  $y_1$  and  $y_2$  are  $n$ -dimensional random fields,  $g$  and  $f$  are continuous (non)linear mappings sufficiently regular, and  $\psi \sim \mathcal{N}(m_\psi, K_\psi)$  and  $\eta \sim \mathcal{N}(m_\eta, K_\eta)$  are random (forcing) vectors that have the same role as defined in (1).

Covariance modeling for GP regression with multiple outputs can be roughly classified into three situations, depending on the amount of information one has about the data source or process. In the first case, the variables (outputs) are known or assumed to be mutually independent, and thus the system can be decoupled and solved separately as two or more unrelated problems. In the second case, we assume that the processes are correlated, but we have no information about the correlation function. In this case, a model can be proposed, or nonparametric inferences can be carried out. In the third situation, we assume that the outputs have a known relationship among them, such as (4) and (5); then the question is how to include this information in the covariance model. The last point forms the scope of this study.

The rest of this paper is organized as follows. We next introduce covariance models and functions that are based on differentiable functions. In Section 2.2 we discuss the implications of using nonlinear relationships and provide high-order closures for the nonlinear models. Analytic auto- and cross-covariance functions are introduced in Section 2.3. Extensive numerical and validation experiments are described in Section 3. We conclude with some final remarks.

## 2. MULTIDIMENSIONAL COVARIANCE MODELS AND FUNCTIONS

In the following section we present models for multidimensional covariance matrices with different closure assumptions. We also introduce analytic forms for covariance functions that fall into the scope of this study.

### 2.1 Multidimensional Covariance Models

Let us denote the mathematical expectation  $E\{y\}$  by  $\bar{y}$  and small perturbations around the expected value by  $\delta y = y - \bar{y}$ . We denote the covariance of two random vectors as  $\text{Cov}(y_i, y_j) = \overline{\delta y_i \delta y_j^T}$  and in more compact form as  $K_{ij}$ . The following lemma introduces a covariance model for a process with two distinct types of outputs that are related through such a function as described in (5),  $y_1 = g(y_2) + \eta$ .

**Lemma 1** (Covariance models for explicit processes). *If two processes  $y_1$  and  $y_2$  satisfy a physical constraint given by (5) with  $g(\cdot) \in C^2$ , and  $\text{Cov}(y_2, y_2) = K_{22}$ , then the covariance matrix formed by the elements of the two vectors satisfies*

$$\begin{aligned} \text{Cov}([y_1^T, y_2^T]^T) &= \begin{bmatrix} LK_{22}L^T + L\text{Cov}(y_2, \eta) + \text{Cov}(\eta, y_2)L^T + \text{Cov}(\eta, \eta) & LK_{22} + \text{Cov}(\eta, y_2) \\ K_{22}L^T + \text{Cov}(y_2, \eta) & K_{22} \end{bmatrix} \\ &+ \mathcal{O}(\delta y_2^3), \end{aligned} \quad (6)$$

where  $L = \partial g / \partial y|_{y=\bar{y}}$  is the Jacobian matrix of  $g$  evaluated at  $E\{y_2\}$ .

*Proof.* The Taylor expansion of  $g(y_2)$  about  $E\{y_2\}$  gives

$$g(y_2) = g(\bar{y}_2) + L\delta y_2 + \frac{1}{2}\delta y_2^T H \delta y_2 + \mathcal{O}(\delta y_2^3),$$

where  $H = \partial^2 g / \partial y^2|_{y=\bar{y}}$ . Take the expectation of (5)

$$\begin{aligned} \bar{y}_1 &= \overline{g(y_2)} + \bar{\eta} = g(\bar{y}_2) + \frac{1}{2}\overline{\delta y_2^T H \delta y_2} + \overline{\mathcal{O}(\delta y_2^3)} + \bar{\eta}, \\ &= g(\bar{y}_2) + \frac{1}{2}\text{tr}(H \text{Cov}(y_2, y_2)) + \overline{\mathcal{O}(\delta y_2^3)} + \bar{\eta}, \end{aligned} \quad (7)$$

where the last line is obtained by using the properties of the trace [ $\text{tr}(A) = \sum A_{ii}$ ] and expectation operators. Equation (7) represents an estimator for the mean process. Subtract (5) from (7), expand  $g(y_2)$ , postmultiply by  $\delta y_2^T$ , and take the expectation on both sides:

$$\overline{\delta y_1 \delta y_2^T} = L \overline{\delta y_2 \delta y_2^T} + \overline{\left( \frac{1}{2} \delta y_2^T H \delta y_2 - \frac{1}{2} \overline{\delta y_2^T H \delta y_2} \right) \delta y_2^T} + \overline{\left( \mathcal{O}(\delta y_2^3) - \overline{\mathcal{O}(\delta y_2^3)} \right) \delta y_2^T} + \overline{\delta \eta \delta y_2^T}.$$

To obtain the cross-covariance block in the right-hand side of (6), we close the system by eliminating terms of  $\mathcal{O}(\delta y_2^3)$ , and obtain

$$\text{Cov}(y_1, y_2) = L \text{Cov}(y_2, y_2) + \text{Cov}(\eta, y_2),$$

or in short  $K_{12} = LK_{22} + K_{\eta 2}$ . To compute  $K_{11}$ , we apply a similar procedure and obtain (6).

Next we have to show that covariance model (6) is an admissible covariance matrix, that is,  $\text{Cov}([y_1^T, y_2^T]^T)$  is symmetric positive definite. A symmetric matrix is positive definite if and only if a subblock and its Schur complement are both positive definite. This can be shown by using minimization of quadratic forms [40] or exploiting properties of determinants [41]. If we pick  $K_{22}$  and its Schur complement  $S = (LK_{22}L^T + LK_{2,\eta} + K_{\eta,2}L^T + K_{\eta\eta}) - (LK_{22} + K_{\eta,2})K_{22}^{-1}(LK_{22} + K_{\eta,2})^T$ , we have that

$$S = K_{\eta\eta} - K_{\eta,2}K_{22}^{-1}K_{\eta,2}^T, \quad (8)$$

which is positive definite because the expression of  $S$  in (8) is also the Schur complement of  $K_{22}$  from

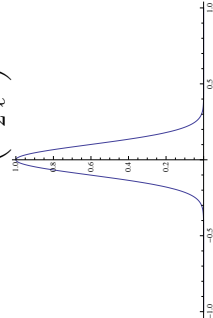
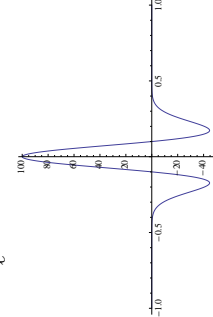
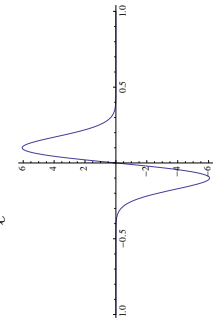
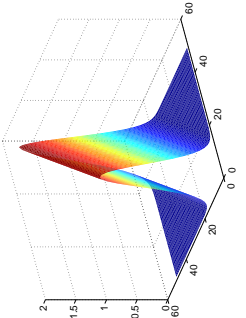
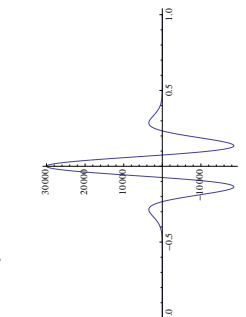
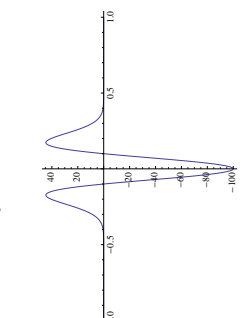
$$\text{Cov}([\eta^T, y_2^T]^T) = \begin{bmatrix} K_{\eta,\eta} & K_{\eta,2} \\ K_{2,\eta} & K_{22} \end{bmatrix}, \quad (9)$$

which is by construction an admissible covariance matrix. This also results from the marginalization property of Gaussian processes.  $\square$

Note that in (6) we write the joint covariance matrix in terms of  $K_{22}$  and noise components, but in practice  $K_{22}$  is not necessarily known. Nonetheless, by using the covariance form introduced in Lemma 1, the inference process fits only the parameters of  $K_{22}$  and noise for all dependent fields. We also remark that if  $g$  is nonlinear in  $y$ ,  $L$  may depend on the mean (value process). The process of ignoring the terms  $\mathcal{O}(\delta y_2^3)$  and of higher order in (7) will be referred to as second-order closure assumption.

The procedure outlined above can also be used to derive analytic closed forms for covariance and cross-covariance functions driven by linear and nonlinear (physics) operators. A few examples for squared exponential and the more general Matérn functions [5] are discussed in Section 2.3 and shown in Fig. 1 for two linear operators [ $g(y) = [\partial/\partial x] y$  and  $g(y) = [\partial^2/\partial x^2] y$ ] and a quadratic one [ $g(y) = y^2$ ].

In the following proposition generalize the result introduced by Lemma 1 to implicit separable systems.

Kernel $K_{22} = \text{Cov}(y_2, y_2)$	Physics	[auto-covariance] $K_{11} = \text{Cov}(y_1, y_1)$ ;	[cross-covariance] $K_{12} = \text{Cov}(y_1, y_2)$ ;
$k(d) = \sigma^2 \exp\left(-\frac{1}{2} \frac{d^2}{\ell^2}\right)$ 	$y_1 = \frac{\partial}{\partial x} y_2$	$\sigma^2 \frac{d^2 - \ell^2}{\ell^4} e^{-\frac{d^2}{2\ell^2}}; [f(\omega) = \ell \sigma \omega^2 e^{-\frac{1}{2} \ell^2 \omega^2}]$  $-K_{22}^{(2)}$	$\sigma^2 \frac{d}{\ell^2} e^{-\frac{d^2}{2\ell^2}}$  $-K_{22}$
	$y_1 = \frac{\partial^2}{\partial x^2} y_2$	$\sigma^2 \frac{d^4 - 6d^2\ell^2 + 3\ell^4}{\ell^8} e^{-\frac{d^2}{2\ell^2}}; [f(\omega) = \ell \sigma \omega^4 e^{-\frac{1}{2} \ell^2 \omega^2}]$  $K_{22}^{(4)}$	$-\sigma^2 \frac{d^2 - \ell^2}{\ell^4} e^{-\frac{d^2}{2\ell^2}}$  $K_{22}^{(2)}$
	$y_1 = y_2^2$	$4 (\mu_{y_2}(x))^2 k(d) + 2 [k(d)]^2$	$2 \mu_{y_2}(x) k(d)$
$k(d) = \sigma^2 \frac{2^{1-\frac{\nu}{2}}}{\Gamma(\nu)} \left(\frac{d\sqrt{\nu}}{\ell}\right)^\nu \times K_\nu\left(\frac{d\sqrt{2\nu}}{\ell}\right)$	$y_1 = \frac{\partial}{\partial x} y_2$	$\sigma^2 \frac{2^{1-\frac{\nu}{2}} \nu}{\ell^3 \Gamma(\nu)} \left(\frac{d\sqrt{\nu}}{\ell}\right)^{\nu-1} \left(\sqrt{2\ell} K_{\nu-1}\left(\frac{d\sqrt{2\nu}}{\ell}\right) - 2d\sqrt{\nu} K_{\nu-2}\left(\frac{d\sqrt{2\nu}}{\ell}\right)\right)$	$\sigma^2 \frac{2^{\frac{3}{2}-\frac{\nu}{2}}}{d \Gamma(\nu)} \left(\frac{d\sqrt{\nu}}{\ell}\right)^{\nu+1} K_{\nu-1}\left(\frac{d\sqrt{2\nu}}{\ell}\right)$
	$y_1 = \frac{\partial^2}{\partial x^2} y_2$	$\sigma^2 \frac{2^{2-\frac{\nu}{2}}}{d^4 \ell^3 \Gamma(\nu)} \left(\frac{d\sqrt{\nu}}{\ell}\right)^{\nu+1} (d\sqrt{\nu} (2d^2\nu + 3\ell^2) K_{\nu-4} \times \left(\frac{d\sqrt{2\nu}}{\ell}\right) + 3\sqrt{2\ell} (\ell^2(\nu-3) - 2d^2\nu) K_{\nu-3} \left(\frac{\sqrt{2}d\sqrt{\nu}}{\ell}\right))$	$-\sigma^2 \frac{2^{1-\frac{\nu}{2}} \nu}{\ell^3 \Gamma(\nu)} \left(\frac{d\sqrt{\nu}}{\ell}\right)^{\nu-1} (2d\sqrt{\nu} K_{\nu-2} \left(\frac{\sqrt{2}d\sqrt{\nu}}{\ell}\right) - \sqrt{2\ell} K_{\nu-1} \left(\frac{\sqrt{2}d\sqrt{\nu}}{\ell}\right))$
	$y_1 = y_2^2$	$4 (\mu_{y_2}(x))^2 k(d) + 2 [k(d)]^2$	$2 \mu_{y_2}(x) k(d)$

**FIG. 1:** Cross- and auto-variance functions of squared exponential and Matérn functions for  $y_1 = (\partial/\partial x)y_2$ ,  $y_1 = (\partial^2/\partial x^2)y_2$ , and  $y_1 = y_2^2$ . Sensible assumptions, such as positivity of the length scale, are considered.

**Proposition 2.1.**

Consider a process driven by the implicit separable system (2). Then under the second-order closure assumptions the block covariance matrix elements satisfy the following simultaneous algebraic equations:

$$\mathbb{K} = \mathbb{L}\mathbb{K}\mathbb{L}^T + \mathbb{L}K_{y\psi} + K_{\psi y}\mathbb{L}^T + K_{\psi\psi}, \quad (\mathbb{K})_{ij} = K_{ij} = \text{Cov}(y_i, y_j), \quad (\mathbb{L})_{ij} = L_{ij} = \frac{\partial f_i}{\partial y_j}. \quad (10)$$

In addition, given  $K_{22}$  and if  $(I - L_{11})$  is invertible, then for the reduced system (4), the following hold:

$$\begin{aligned} K_{11} = & L_{11}K_{11}L_{11}^T + L_{12} \left( (I - L_{11})^{-1} (L_{12}K_{22} + K_{\psi_{1,2}}) \right)^T L_{11}^T \\ & + L_{11} \left( (I - L_{11})^{-1} (L_{12}K_{22} + K_{\psi_{1,2}}) \right) L_{12}^T + L_{12}K_{22}L_{12}^T \\ & + K_{\psi_{1,1}}L_{11}^T + K_{\psi_{1,2}}L_{12}^T + L_{11}K_{1,\psi_1} + L_{12}K_{2,\psi_1} + K_{\psi_1\psi_1}, \end{aligned} \quad (11)$$

$$K_{12} = (I - L_{11})^{-1} (L_{12}K_{22} + K_{\psi_{1,2}}). \quad (12)$$

*Proof.* We use Lemma 1 and the chain rule on (2). We illustrate the calculations on the system with two vector components (4). The following relations are obtained:

$$\begin{aligned} f_i(y_1, y_2) &= f_i(\bar{y}_1, \bar{y}_2) + L_{i1}\delta y_1 + L_{i2}\delta y_2 + \mathcal{O}(\delta y^2), \\ \bar{y}_i &= f_i(\bar{y}_1, \bar{y}_2) + \mathcal{O}(\bar{\delta y}^2) + \bar{\psi}_i, \\ \delta y_i &= L_{i1}\delta y_1 + L_{i2}\delta y_2 + \delta\psi_i, \end{aligned}$$

where  $i = 1, 2$  and the second line is obtained from the first and (2). Then under the second-order closure assumptions and by using the same procedure in Lemma 1, one obtains

$$K_{12} = \overline{\delta y_1 \delta y_2^T} = L_{11}K_{12} + L_{12}K_{22} + K_{\psi_{1,2}}.$$

If  $(I - L_{11})$  is invertible, then one obtains (12). It can be shown that  $K_{11}$  satisfies

$$\begin{aligned} \overline{\delta y_1 \delta y_1^T} = & L_{11}K_{11}L_{11}^T + L_{12}K_{21}L_{11}^T + L_{11}K_{12}L_{12}^T + L_{12}K_{22}L_{12}^T \\ & + K_{\psi_{1,1}}L_{11}^T + K_{\psi_{1,2}}L_{12}^T + L_{11}K_{1,\psi_1} + L_{12}K_{2,\psi_1} + K_{\psi_1\psi_1}. \end{aligned}$$

Then the substitution of  $K_{12}$  in (12) yields (11). The terms can be collected and expressed as in (10). Relation (11) is obtained by eliminating  $K_{12}$  from (10).  $\square$

In general, cross covariances involving  $y_2$  and the forcing  $\psi$  are difficult to specify. Nevertheless, we consider it important to preserve such terms in order to have a complete representation of the problem. In our numerical experiments, however, we will treat these terms as zero.

The procedure in Proposition 2.1 can also be interpreted as an extension of the Delta method procedure [42, 43]. We also note the agreement between (10) and the results presented in [18], in which a particular hyperbolic PDE-driven process is explored. Thus far we have assumed that the physical constraint is exactly or well approximated by the first derivative, implying that the physics is approximately well represented by its linearization. If this is not the case, then higher-moment closures for a Gaussian processes can be explored [44]. Alternatively, closure for non-Gaussian distributions is discussed in [45–47].

**2.2 High-Order Closures**

Closures with higher-order moments can lead to more accurate models for nonlinear processes. We now develop the covariance model with third-order closure of the error truncation terms for problem (5).

To use high-order expansions, we will use tensor algebra (in Cartesian coordinates) with the following conventions. The Hessian tensor of  $g(\mathbf{y})$  is given by a rank-three tensor  $H_{ijk} = [\partial^2 g_j(\mathbf{y})] / \partial y_i \partial y_k$ . The transpose of a tensor

is obtained by permuting the first and last indices:  $(H_{ijk})^T = H_{kji}$ . The trace of a rank-three tensor is a tensor contraction to a (rank-one tensor or a) vector defined as  $\text{tr}(H) = \{\sum_{ik} H_{ijk} \delta_{ik}\}_j$ . The product of a rank-three tensor with a matrix is defined in the usual dot product sense, resulting in a rank-three tensor. By using these conventions we can represent the algebraic relations in this section using rules that mimic the scalar-valued function case.

**Proposition 2.2.**

Consider a process driven by (5). Then under the third-order closure assumptions, its covariance matrix takes the following form:

$$\text{Cov}([y_1^T, y_2^T]^T) = \begin{bmatrix} K_{11} & LK_{22} + \text{Cov}(\eta, y_2) \\ K_{22}L^T + \text{Cov}(y_2, \eta) & K_{22} \end{bmatrix} + \mathcal{O}(\delta y_2^3), \quad (13)$$

where  $L$  is the Jacobian matrix,  $H$  is the Hessian tensor corresponding to  $g$  evaluated at  $E\{y_2\}$ , and

$$K_{11} = K_{11}^I + \frac{1}{4} \overline{\delta y_2^T H \delta y_2 \delta y_2^T H^T \delta y_2} - \frac{1}{4} \text{tr}(HK_{22}) \text{tr}(HK_{22})^T + \frac{1}{2} \overline{\delta y_2^T H \delta y_2 \eta^T} + \frac{1}{2} \overline{\delta \eta \delta y_2^T H^T \delta y_2}, \quad (14)$$

where  $K_{11}^I$  is the corresponding term using second-order closure assumptions.

*Proof.* The Hessian is a symmetric operator in the sense defined above ( $H = H^T$ ) because we assume that  $g$  is a smooth function. We therefore have

$$\delta y_1 = L \delta y_2 + \left( \frac{1}{2} \delta y_2^T H \delta y_2 - \frac{1}{2} \overline{\delta y_2^T H \delta y_2} \right) + \left( \mathcal{O}(\delta y_2^3) - \overline{\mathcal{O}(\delta y_2^3)} \right) + \delta \eta,$$

$$\delta y_1 \delta y_2^T = L \delta y_2 \delta y_2^T + \frac{1}{2} \delta y_2^T H \delta y_2 \delta y_2^T - \frac{1}{2} \text{tr}(HK_{22}) \delta y_2^T + \delta \eta \delta y_2^T,$$

$$K_{12} = LK_{22} + \frac{1}{2} \overline{\delta y_2^T H \delta y_2 \delta y_2^T} + K_{\eta 2} = K_{12}^I,$$

where  $\overline{\delta y_2^T H \delta y_2 \delta y_2^T}$  has only third-order central moments and is therefore zero (for normals), and  $K_{12}^I$  is the second-order approximation of the cross covariance.

Block  $K_{11}$  follows as

$$\begin{aligned} K_{11} = & K_{11}^I + \frac{1}{2} \overline{L \delta y_2 \delta y_2^T H^T \delta y_2} + \frac{1}{2} \overline{\delta y_2^T H \delta y_2 \delta y_2^T} L^T + \frac{1}{4} \overline{\delta y_2^T H \delta y_2 \delta y_2^T H^T \delta y_2} \\ & - \frac{1}{4} \text{tr}(HK_{22}) \text{tr}(HK_{22})^T + \frac{1}{2} \overline{\delta y_2^T H \delta y_2 \eta^T} - \frac{1}{4} \text{tr}(HK_{22}) \text{tr}(HK_{22})^T \\ & + \frac{1}{4} \text{tr}(HK_{22}) \text{tr}(HK_{22})^T + \frac{1}{2} \overline{\delta \eta \delta y_2^T H^T \delta y_2}, \end{aligned}$$

where  $K_{11}^I$  is given by (6). Then  $K_{11}$  reduces to

$$\begin{aligned} K_{11} = & K_{11}^I + \frac{1}{4} \overline{\delta y_2^T H \delta y_2 \delta y_2^T H^T \delta y_2} - \frac{1}{4} \text{tr}(HK_{22}) \text{tr}(HK_{22})^T \\ & + \frac{1}{2} \overline{\delta y_2^T H \delta y_2 \eta^T} + \frac{1}{2} \overline{\delta \eta \delta y_2^T H^T \delta y_2}. \end{aligned} \quad (15)$$

□

We see that relaxing the closure assumptions does not complicate the structure of the entire covariance matrix; however, block  $K_{11}$  appears to present computational difficulties. Nonetheless, by using the relation among moments in normal distributions, we observe that the quartic term in (15), which is potentially the most difficult term to calculate, can be factorized in terms of entries in  $K_{22}$  [48]:

$$\begin{aligned}\overline{(\delta y_2(i))^4} &= 3K_{22}^2(i, i) \\ \overline{(\delta y_2(i))^2(\delta y_2(j))^2} &= K_{22}(i, i) \times K_{22}(j, j) + 2K_{22}^2(i, j).\end{aligned}$$

Such models are exact for quadratic processes such as the derivatives occurring in Burgers equations; that is,  $\partial y^2/\partial x$  or  $y \partial y/\partial x$  for solution  $y(x)$ . For more complicated problems, truncated model (13) represents an approximation of the covariance matrix by accounting for a one-way action of high-order moments. In Section 3.2 we provide a simple but illustrative example in which the high-order closure assumptions lead to a more robust matrix structure.

## 2.3 Auto- and Cross-Covariance Functions

We now focus on the analytic forms of the covariance functions that are used to generate the covariance matrices discussed above. The analytical covariance functions provide distinct theoretical and practical advantages by allowing the design of grid-independent covariance structures and facilitating a rigorous asymptotic analysis. By using the same calculation procedures we arrive at several covariance models. However, because the analytic forms depend on the covariance function of the independent process (i.e.,  $K_{22}$ ,  $y_2$ ) and the expression that relates it to the dependent one (i.e.,  $g$ ), we limit our results to a few processes. In Fig. 1 we present the auto- and cross-covariance functions for processes driven by the following:  $y_1 = (\partial/\partial x)y_2$ ,  $y_1 = (\partial^2/\partial x^2)y_2$ , and  $y_1 = y_2^2$ ; these functions may correspond to processes such as pressure-wind, diffusion, or force-acceleration, respectively. In every case  $y_2$  is modeled either by squared exponential or by Matérn covariance functions. These results can be obtained by using the intermediate steps in the proof of Lemma 1 or can be derived from their characteristic functions by differentiating the kernels in the Fourier space; see [5].

Another strategy to arrive at the same results is the following. We assume that the  $y_2$  process is weakly stationary and sufficiently mean-square differentiable (in the Matérn case) and all metrics and multiplicative factors are positive [5]. Then, with the standard kriging notation, we recover the results presented in [5] for the mean square differentiable process  $Z(x_1, x_2)$  and for which one obtains the auto-covariance function of the differentiated process as  $k_{\dot{Z}}(d) = -k''(d)$ , where  $d = |x_1 - x_2|$  (absolute value) and  $k(d) = \text{Cov}(Z, Z)$ . A similar strategy, but in a different coordinate system, is described in [7, 17, 26]. In this simple case, the Fourier transform of the covariance function can also be used to compute the resulting auto- and cross-correlation functions; however, some functionals (processes and kernels) may lead to difficult calculations and the contribution of higher-order moments would be more difficult to assess.

The linear processes lead to simple auto- and cross-covariance functions. In the nonlinear case the kernel takes a parametric form that depends on the mean value process, which comes as no surprise. In the top part of Fig. 1 we illustrate the square exponential covariance function and the generated auto- and cross covariances. We also present a graphical illustration of the kernels for scalar variables with some fixed coefficients, and a three-dimensional (matrix) representation of  $K_{22}$ , with variance equal to 2. The auto-covariance functions need to be positive definite, and for reference in the ‘‘Gaussian’’ kernel case we also give their respective Fourier transform,  $f(\omega)$ . The lower part of Fig. 1 presents the same results for the Matérn functions. A numerical validation of the covariance functions described in Fig. 1 is given in Section 3.2.

## 3. NUMERICAL EXPERIMENTS AND VALIDATION

In this section we provide a numerical illustration of the theoretical considerations introduced in the preceding sections. We begin with a numerical validation of the new covariance models and present a few nonstationarity considerations. We continue with applying the theoretical considerations introduced in this study to one-dimensional spatial interpolation experiments. Here we perform several Gaussian process linear regression experiments under a controlled setting. In the section that follows we investigate linear regression using the covariance models on a two-dimensional problem that is based on a real data set.

In all the regression experiments the inference process carried with the covariance models introduced in this study is compared with an independent fit of the different quantities. This approach is intended to expose the efficiency gains when the covariance structure is properly specified.



### 3.1 One-Dimensional Model Process

We first introduce a problem inspired by the geostrophic balance (atmosphere in equilibrium) [38], which relates wind speed and pressure after a series of approximations: wind  $\propto \partial/\partial x$  pressure, and therefore  $g(\cdot) = \partial/\partial x$ . To give more physical intuition, we use  $u$  for wind and  $p$  for pressure. For this model we evaluate the following example:

$$\begin{aligned} p &\sim \mathcal{N}(0, C_{se}), \quad x \in [-1, 1], \quad \Delta x = 2/100, \\ u &= \frac{\partial p}{\partial x}, \quad g(p) = \frac{\partial}{\partial x} p, \end{aligned} \quad (16)$$

where  $C_{se}(i, j) = \sigma^2 \exp\{-(1/2)[x(i) - x(j)]^2/\ell^2\}$ . For our synthetic example we take  $\ell = \ell_2 = 7\Delta x/\sqrt{2}$ . We also consider diffusion,  $g(y) = (\partial^2/\partial x^2)y$ , as well as a nonlinear operator,  $g(y) = y^2$ , which are both in  $C^2$ .

For regression problems we consider the following bivariate one-dimensional model:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} u(x) \\ p(x) \end{bmatrix}, \quad x \in [-1, 1], \quad x_i = 2\Delta x i - 1, \quad i = 1, \dots, 100, \quad \Delta x = 2/100, \quad (17)$$

$$y_1 = u(x) = g(p(x)) = \alpha \frac{\partial}{\partial x} p(x) + \eta, \quad \eta \sim \mathcal{N}(0, C_{mat, \nu=5/2}([\ell_\eta, \sigma_\eta^2])), \quad \text{Cov}(y_2, \eta) \equiv 0, \quad (18)$$

$$y_2 = p(x) \sim \mathcal{N}(0, C_{se}([\ell_2, \sigma_2^2])), \quad (19)$$

where  $u$  represents the wind field,  $p$  takes the place of the pressure, and  $C_{mat, \nu=5/2}$  is the Matérn covariance function with the corresponding smoothness. This process is observed at various grid points  $f_i$  with an additive normal noise  $\varepsilon$ , that is,

$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = H_{\mathbf{y}} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_{n,i}^2 I), \quad (20)$$

where  $H_{\mathbf{y}}$  represents an observation operator that picks values corresponding to selected grid points. In this case the Jacobian matrix corresponds to a differential operator,  $L = \alpha(\partial/\partial x)$ .

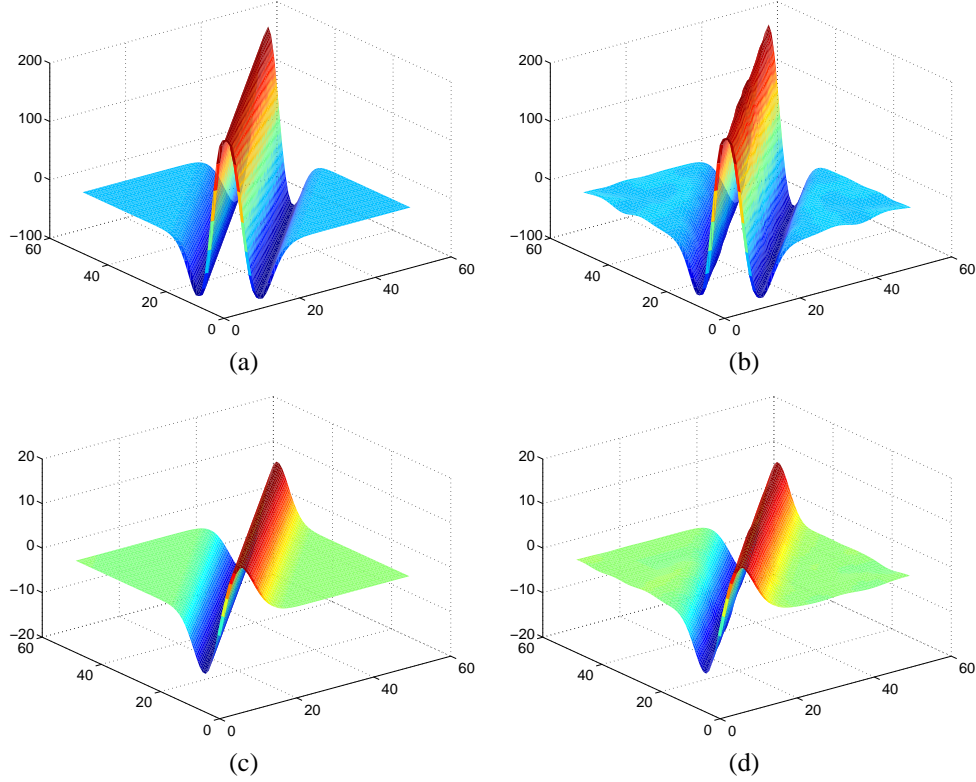
We note that the particular form of processes described by (16) and (17)–(20) is rather general and common in practice, where one may find observations that represent the state and their derivatives (in space) also called tendencies in the atmospheric sciences [17, 25, 26, 38].

### 3.2 Validation: Examples of Covariance Models

We begin by considering the set of governing equations presented in Fig. 1 and calculate the cross- and auto-covariance functions for each case for the two-dimensional kernel  $K_{22}$  represented therein. The analytical results are contrasted with sample-based covariance estimation.

In Fig. 2 we show the auto-covariance ( $K_{11}$ ) and cross-covariance ( $K_{12}$ ) obtained from 10,000 samples as well as the analytical model that corresponds to (17)–(20). Observe the almost perfect resemblance as well as the equivalence among these realizations and the analytic kernels described in the upper segment of Fig. 1. In Fig. 3 we illustrate the sample-based and analytic covariance matrices for the diffusion process. The same number of samples was used as before. The relatively small difference between the two approximations can be attributed to sampling noise; that is, increasing the number of samples reduced the size of the difference. Of course, because the model is linear, we expect the formula to be exact. The  $L$  operator is obtained from a finite difference approximation of  $g$ ; the two approximations considered here and throughout this study are  $y'(x) \approx [y(x_{i+1}) - y(x_i)]/\Delta x$  and  $y''(x) \approx [y(x_{i+1}) - 2y(x_i) + y(x_{i-1})]/\Delta x^2$ . We note that in this case samples from the exact distribution can be drawn and it is not necessary to use a finite difference approximation; however, we attempt to replicate a setting in which the samples resulting from other processes are not as easily obtained.

We now discuss results obtained by using the covariance matrix model with high-order closure assumptions as given by Proposition 2.1 and the one based on the second-order closure as described in Lemma 1 that correspond to the quadratic process  $y_1 = y_2^2 + \eta$ . In this case [as explained by (14)]  $K_{11}$  differs between the two covariance models;



**FIG. 2:** Auto-covariance and cross-covariance matrices obtained using a first-order derivative physics model:  $y_1 = u(x) = g(p(x)) = \alpha(\partial/\partial x)p + \eta$ ,  $y_2 = p \sim \mathcal{N}(0, C_{se}([\ell_2, \sigma_2^2]))$  and  $\eta \sim \mathcal{N}(0, C_{mat, \nu=5/2}([\ell_\eta, \sigma_\eta^2]))$ . The results in (a, c) are obtained by using the physics-based approach, that is Lemma 1, and the results in (b, d) are obtained by sample approximation based on 10,000 simulations.

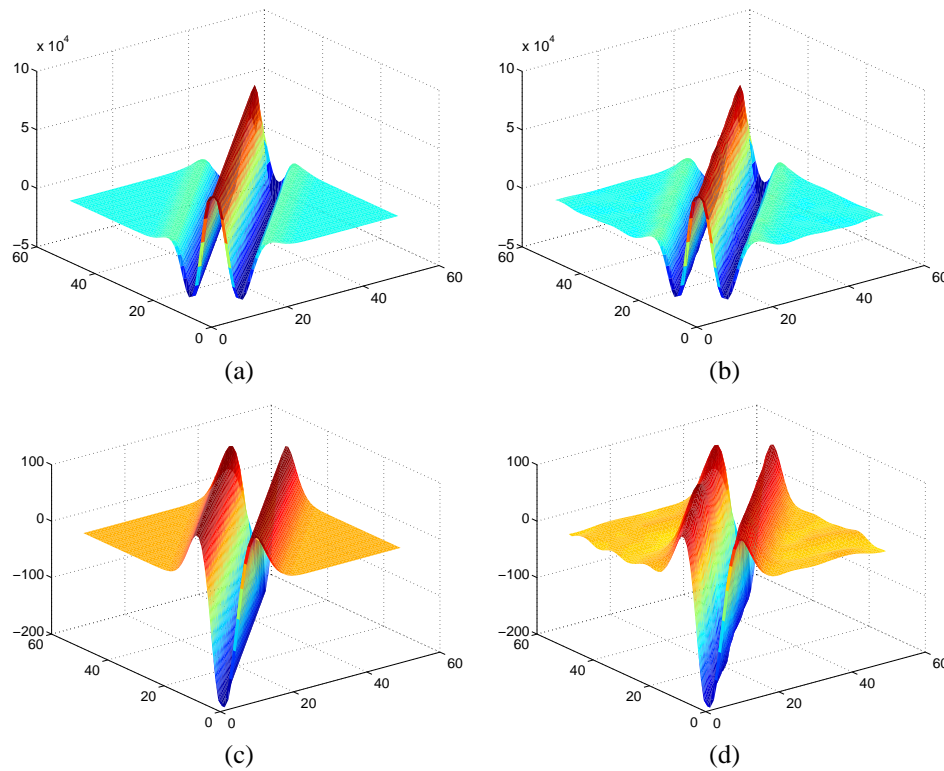
however,  $K_{12}$  is practically unchanged. We therefore focus on  $K_{11}$ , and in Fig. 4 we illustrate the absolute errors in the covariance matrix entries between the quadratic and cubic covariance matrix models and two sample-based approximations using 10,000 and 50,000 samples. The error levels estimated by using 10,000 samples are relatively similar between the two matrix models; however, we note that when increasing the number of samples used for the error estimate, the quadratic model shows significantly smaller errors. The latter indicates that the high-order closure provides a more robust covariance matrix approximation in this case.

This set of three experiments corresponds to the particular setting described in Fig. 1 and illustrates the theoretical statements presented in Propositions 2.1 and 2.2.

### 3.3 Physics-Induced Nonstationarity

As discussed in the introductory part of this study, nonstationary covariance models are generally difficult to construct, and typical strategies rely on changing smoothing properties or correlation distances in different directions. We now briefly discuss nonstationary models induced in the covariance structure through procedures introduced in this study. For brevity we present a one-dimensional case with one variable; nonetheless, this strategy can be extended to multiple variables and dimensions and is applicable on all the following examples. A situation in which no stationary models are suitable occurs when the domain geometry (or topology) is not uniform, such as shallow-water approximations and subsurface flows.

We consider the first-order differential model in space  $y_1 = a(x)y_2(x)'$ , where  $a(x)$ , for instance, represents a constraint or a mapping from an irregular grid to a regular one. From Lemma 1 it follows that  $K_{11} = LK_{22}L^T$ , where



**FIG. 3:** Auto-covariance and cross-covariance matrices obtained by (a, c) using Lemma 1, and a second-order derivative (Laplace operator) model:  $y_1 = u(x) = g(p(x)) = \alpha(\partial^2/\partial x^2)p + \eta$ , and (b, d) sample approximation based on 10,000 simulations.

$L$  is the spatial differential operator discretized as  $y(x_i) = a(x_i) * [y(x_i) - y(x_{i-1})]/\Delta x$ ; alternatively  $a(x)$ , can take the place of a fixed  $\Delta x$ .

In Fig. 5(a) we illustrate samples drawn from distributions induced from the differential operator with spatial variability. In one case we consider  $a(x)$  to be a fixed value; in the other in an *ad hoc* manner we choose  $a(x)$  to be

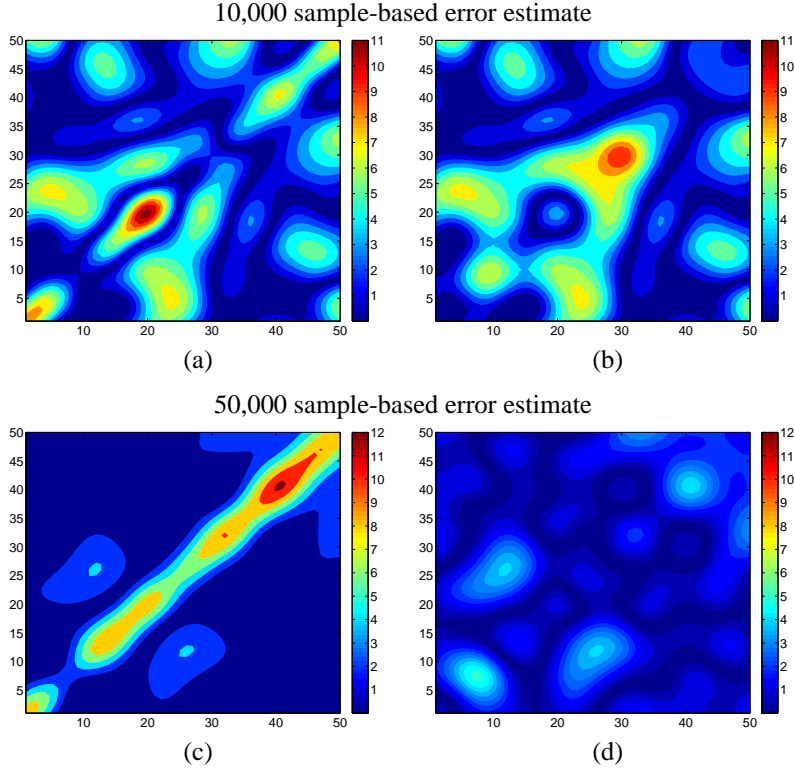
$$a(x) = \frac{3}{2} + \sin\left(1 + \frac{x\pi}{N}\right)^2 + \left(\frac{1}{2} + \left(\frac{x}{N}\right)^3\right) \cos\left(1 + \frac{15x\pi}{N}\right), \quad (21)$$

where  $N$  is the number of grid points. The induced covariance  $K_{11}$ , shown in Fig. 5(b), and ten samples drawn from  $K_{11}$  with  $a(x) = 1$  and with  $a(x)$  defined by (21) are contrasted in Figs. 5(c) and 5(d), respectively. We can also interpret  $a(x)$  to be the reciprocal of the fixed grid spacing  $\Delta x$ . The samples in Fig. 5(c) have a noticeable spatial structure induced in terms of length scale and variance through the use of  $a(x)$ . In particular, note the sample “clamping” that takes place around grid point 90 and the relatively wavy structure of the weighting function that can be seen in the sample behavior.

Note also that the nonstationarity is introduced directly in the discretization of the physical process and therefore provides a more consistent structure than does treating the physics and coordinate transformations separately.

### 3.4 GP Regression Using Physics-based Covariance Models

We compute the joint distribution and posterior applied in a GP regression problem with multiple outputs. The central example is the geostrophic wind (17)–(20); nonetheless, we also illustrate GP regression results for systems governed by other processes. Henceforth, quantities subscript  $*$  ( $\circ_*$ ) represent predictions.



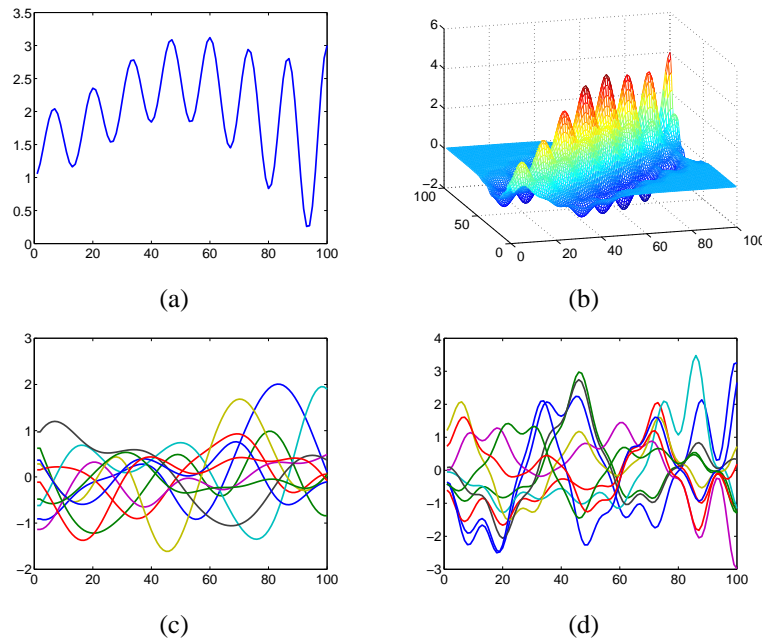
**FIG. 4:** Pointwise errors in the auto-covariance ( $K_{11}$ ) for a quadratic (physics) model:  $y_1 = y_2^2 + \eta$ ,  $y_2 \sim \mathcal{N}(6, C_{se}([\ell_2, \sigma_2^2]))$  and  $\eta \sim \mathcal{N}(0, C_{mat, \nu=5/2}([\ell_\eta, \sigma_\eta^2]))$ , by using (a, b) 10,000 samples and (c, d) 50,000 samples. We show the error estimates for two covariance models that correspond to (a, c) a quadratic closure assumption (6) and (b, d) a cubic closure assumption (10) of the physical process. The error is calculated as the absolute pointwise difference between the corresponding sample-based covariance estimate and the physics-based one. We note that improving the accuracy of the sample error estimate yields a significantly reduced error in the covariance matrix entries for the cubically truncated model.

### 3.4.1 Gaussian Process Regression

In this section we consider  $\text{Cov}(y_2, \eta) = 0$  and make predictions for  $y_1(x_*) = y_{*1}$  and  $y_2(x_*) = y_{*2}$ . The joint distribution corresponding to problem (17)–(20) is given by Lemma 1:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_{*1} \\ y_{*2} \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} g(\overline{y_2}) + (1/2)\text{tr}(HK_{22}) \\ \overline{y_2} \\ g(\overline{y_{*2}}) + (1/2)\text{tr}(HK_{*2,*2}) \\ \overline{y_{*2}} \end{bmatrix}, \begin{bmatrix} LK_{22}L^T + K_{\eta\eta} & LK_{22} & LK_{2,*2}L^T + K_{\eta,*\eta} & LK_{*2,*2} \\ K_{22}L^T & K_{22} & K_{2,*2}L^T & K_{*2,*2} \\ LK_{*2,2}L^T + K_{*\eta,\eta} & LK_{*2,2} & LK_{*2,*2}L^T + K_{*\eta,*\eta} & LK_{*2,*2} \\ K_{*2,*2}L^T & K_{*2,*2}^T & K_{*2,*2}L^T & K_{*2,*2} \end{bmatrix} + \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \right), \quad (22)$$

$$\Sigma = \begin{bmatrix} K_{\varepsilon_1 \varepsilon_1} & 0 \\ 0 & K_{\varepsilon_2 \varepsilon_2} \end{bmatrix} = \begin{bmatrix} \sigma_{n,1}^2 I & 0 \\ 0 & \sigma_{n,2}^2 I \end{bmatrix},$$



**FIG. 5:** Nonstationary covariance structure. In (a) we plot (21) and in (b)  $K_{11}$  obtained with (21). In (c) and (d) we show ten samples from  $K_{11}$  obtained with  $a(x) = 1$  and with (21), respectively.

where matrix  $\Sigma$  represents the noise in observations.

The predictive distribution is then obtained as a normal distribution with expectation and covariance matrix given by [3]

$$\begin{aligned} \overline{\mathbf{y}_* | \mathbf{X}, \mathbf{X}_*, \mathbf{y}} &= \mathbf{m}(\mathbf{X}_*) + \mathbf{K}_{21} (\mathbf{K}_{11} + \Sigma)^{-1} [\mathbf{y} - \mathbf{m}(\mathbf{X})], \\ \text{Cov}(\mathbf{y}_* | \mathbf{X}, \mathbf{X}_*, \mathbf{y}) &= \mathbf{K}_{22} - \mathbf{K}_{21} (\mathbf{K}_{11} + \Sigma)^{-1} \mathbf{K}_{12}, \end{aligned}$$

where  $\mathbf{X}$  represents the covariates,  $\mathbf{y}$  the responses,  $\mathbf{y}_*$  the corresponding predictions, and  $\mathbf{K}$  is the covariance matrix in (22) with blocks  $\mathbf{K}_{11}$ ,  $\mathbf{K}_{12}$ ,  $\mathbf{K}_{21}$ ,  $\mathbf{K}_{22}$ . The covariance matrix depends on several parameters, also known as hyperparameters. To fit the hyperparameters in the covariance model, we use a Newton-based strategy to maximize the marginal log-likelihood expression (or evidence) [3]:

$$\log [\mathcal{P}(\mathbf{y} | \mathbf{X}, \theta)] = -\frac{1}{2} [\mathbf{y} - \mathbf{m}(\mathbf{X})]^T (\mathbf{K}_{11} + \Sigma)^{-1} [\mathbf{y} - \mathbf{m}(\mathbf{X})] - \frac{1}{2} \log |\mathbf{K}_{11} + \Sigma| - \frac{n_2}{2} \log(2\pi),$$

with its gradient given by

$$\begin{aligned} \frac{\partial}{\partial \theta_j} \log \mathcal{P}(\mathbf{y} | \mathbf{X}, \theta) &= \frac{1}{2} (\mathbf{y} - \mathbf{m}(\mathbf{X}))^T \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_j} \mathbf{K}^{-1} (\mathbf{y} - \mathbf{m}(\mathbf{X})) - \frac{1}{2} \text{tr} \left( \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_j} \right), \\ &= \frac{1}{2} \text{tr} \left( (\alpha \alpha^T - \mathbf{K}^{-1}) \frac{\partial \mathbf{K}}{\partial \theta_j} \right); \quad \alpha = \mathbf{K}^{-1} [\mathbf{y} - \mathbf{m}(\mathbf{X})]. \end{aligned} \quad (23)$$

We consider a more general framework that includes observational operators. We define two mappings from the observation space to the prediction space,  $\mathcal{H}_{1,*1} : \mathbb{R}^{n_{*1}} \rightarrow \mathbb{R}^{n_1}$  and  $\mathcal{H}_{2,*2} : \mathbb{R}^{n_{*2}} \rightarrow \mathbb{R}^{n_2}$ , as well as their linearizations  $H_{1,*1}$  and  $H_{2,*2}$ , respectively. Let us also consider  $n_{*1} \gg n_1$  and  $n_{*2} \gg n_2$ , which correspond to having sparse observations relative to predictions. For instance, if  $x_1 = [1, 2, 3]^T$  and  $x_{*1} = [2, 3]^T$ , then  $H_{1,*1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^T$ . We also consider the Jacobian matrix  $L(= L_{*1,*2})$  to be a well-defined mapping  $L : \mathbb{R}^{n_{*2}} \rightarrow \mathbb{R}^{n_{*1}}$

introduced in Lemma 1 as well as its projection to the observation space  $L_{1,*2} = H_{1,*1}L_{*1,*2}$ . The elements in the joint distribution can be computed as follows:

$$\mathbf{K}_{11} = \begin{bmatrix} L_{1,*2}K_{*2,*2}L_{1,*2}^T + K_{\eta\eta} & L_{1,*2}K_{*2,*2}H_{2,*2}^T \\ H_{2,*2}K_{*2,*2}L_{1,*2}^T & K_{2,2} \end{bmatrix},$$

$$\mathbf{K}_{12} = \begin{bmatrix} L_{1,*2}K_{*2,*2}L^T + K_{\eta,*\eta} & L_{1,*2}K_{*2,*2} \\ H_{2,*2}K_{*2,*2}L^T & K_{2,*2} \end{bmatrix}, \quad \mathbf{K}_{21} = \mathbf{K}_{12}^T,$$

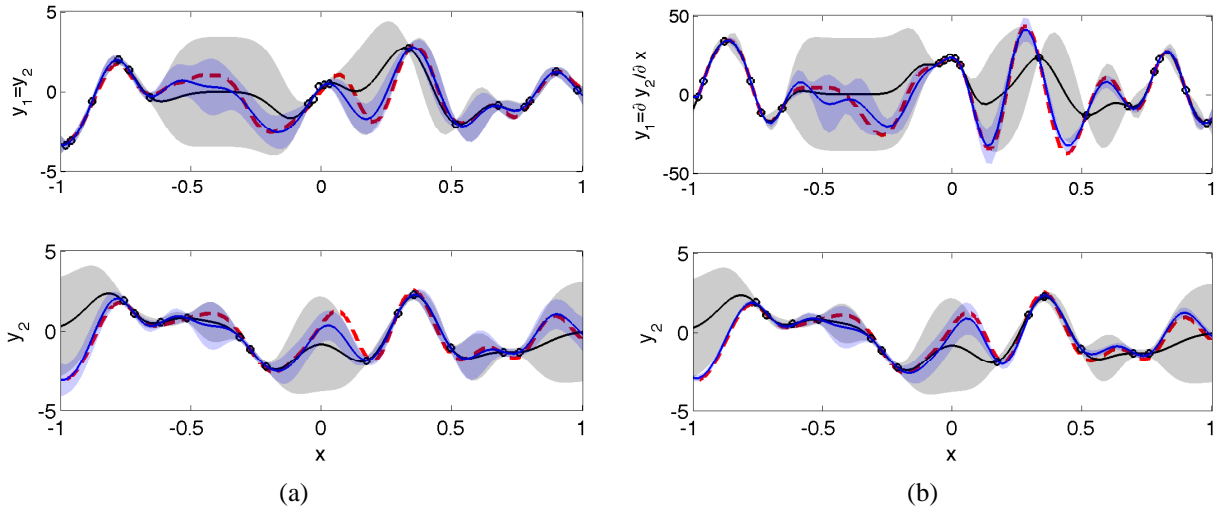
$$\mathbf{K}_{22} = \begin{bmatrix} L^TK_{*2,*2}L + K_{* \eta, * \eta} & L^TK_{*2,*2} \\ (LK_{*2,*2})^T & K_{*2,*2} \end{bmatrix}.$$

We consider the following hyperparameters for problem (17)–(20):  $\theta = \{\ell_2, \ell_\eta, \sigma_2^2, \sigma_\eta^2, \sigma_{n,1}^2, \sigma_{n,2}^2\}$ . These correspond to  $K_{22}$  (also to  $K_{*2,*2}$ ),  $K_{\eta\eta}$  (also to  $K_{* \eta, * \eta}$ ), and  $\Sigma$ .

In the following sections we present numerical results for Gaussian process regression experiments. We use the setup described in Section 3.1; however, we consider different processes  $g$  of increasing complexity. In all experiments we consider the noise in the data to be  $\sigma_n^2 = 0.2$  and the noise in the physical process to be  $\sigma_\eta^2 = 0.1$ .

### 3.4.2 Simple Linear Process

In Fig. 6(a) we present the independent and joint fit of two Gaussian processes for the linear problem  $y_1 = y_2 + \eta$ ,  $y_2 \sim \mathcal{N}(0, C_{se}([\ell_2, \sigma_2^2]))$ , and  $\eta \sim \mathcal{N}(0, C_{mat, \nu=5/2}([\ell_\eta, \sigma_\eta^2]))$ . This means that  $y_1$  is a noisy representation of  $y_2$ . The two processes are on the same grid (“x”). The GP model with the joint fit collects observational information from both covariates when computing the posterior and therefore gives a better prediction. This is not surprising because the regression on the joint process is identical with a regression performed on a single variable with the complete set of observations.



**FIG. 6:** Independent fit of two Gaussian processes (gray) and dependent fit for two models: (a) for a simple linear model:  $y_1 = y_2 + \eta$ ,  $y_2 \sim \mathcal{N}(0, C_{se}([\ell_2, \sigma_2^2]))$ , and  $\eta \sim \mathcal{N}(0, C_{mat, \nu=5/2}([\ell_\eta, \sigma_\eta^2]))$  and (b) a differential model (first-order approximation):  $y_1 = u(x) = g[p(x)] = \alpha(\partial/\partial x)p + \eta$ . The dashed line represents the “truth” with noisy observations denoted by circles. The solid dark line represents the independent fit and the gray shade the point variance. The blue solid line with blue shade represents the dependent fit.

### 3.4.3 First-Order Differential and Laplace Operators

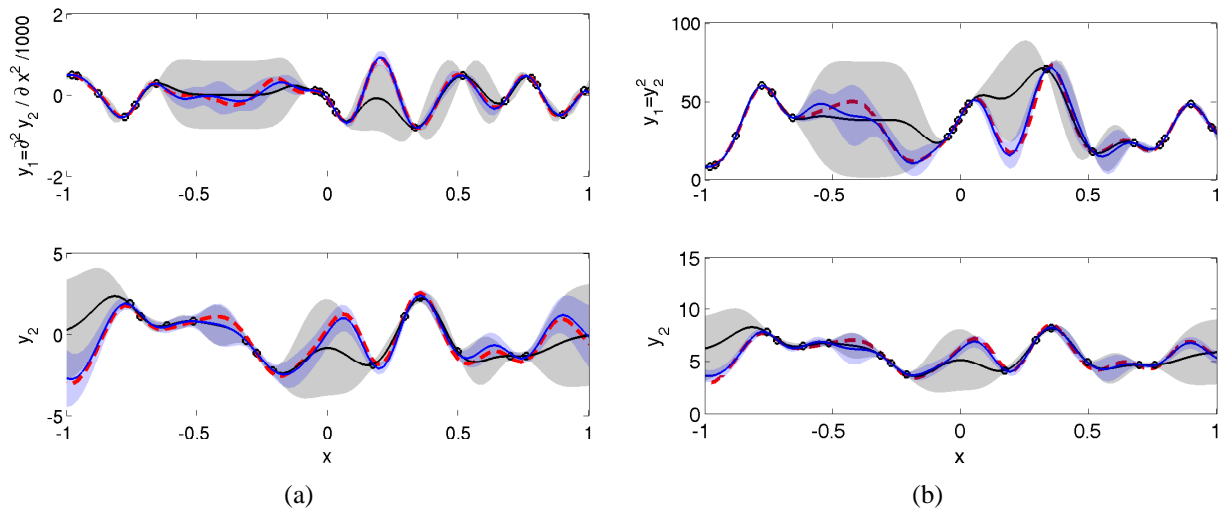
In Fig. 6(b) we show the same experiment but replace the physical process with a first-order differential operator in space, which is approximated by finite differences. We again see the more accurate agreement between prediction and the “true” value when using the joint model. In Fig. 7(a) we present the results for a Laplace operator with the same conclusions. In both cases we note the near-perfect fit on the left boundary of process  $y_2$  that *has no observations*. This is a direct result of information transfer from observations in  $y_1$  present at that location.

### 3.4.4 Quadratic Process

In Fig. 7(b) we illustrate the results for a quadratic “physics” process:  $y_1 = y_2^2 + \eta$ ,  $y_2 \sim \mathcal{N}(6, C_{se}([\ell_2, \sigma_2^2]))$  and  $\eta \sim \mathcal{N}(0, C_{mat, \nu=5/2}([\ell_\eta, \sigma_\eta^2]))$ . The mean process is set to 6 in order to limit the observability issues. For this case we use the high-order closure covariance model introduced in Section 2.2. To illustrate quantitatively the difference between using a second-order closure as assumed in lemma 1 and the high-order closures assumed in proposition 2.2, we compare the root mean square of the error (RMSE) in the prediction obtained using these two models. When using a linear covariance model (lemma 1) the RMSE of  $y_1$  is 3.40, whereas the high-order model (proposition 2.2) used in prediction yields an RMSE of 2.83. Of course  $y_2$  has less improvement, 0.31 from 0.29, because the auto-covariance model is exact; furthermore, an independent fit gives 11.10 for  $y_1$  and 1.12 for  $y_2$ . We argue that these results are reasonable given the nature of this experiment, that is, the fact that the normality assumptions are no longer optimal; however, we note that they are in agreement with our theoretical expectations through an overestimation of the auto-covariance in the quadratic closure model, as indicated in Fig. 4.

## 3.5 Large-Scale Numerical Experiment

In the following section we apply the covariance models described in this paper to an inference problem for geostrophic winds based on data resulting from real regional numerical weather prediction systems. In this case we have two-dimensional fields. A similar experimental setting was presented in [17, 26], and because of its relevancy we choose the same type of problem. Nonetheless, by taking a consistent and systematic approach at constructing our models we



**FIG. 7:** Independent fit of two Gaussian processes (gray) and dependent fit for (a) a differential model - Laplace operator:  $y_1 = u(x) = g[p(x)] = \alpha(\partial^2/\partial x^2)p + \eta$ ,  $y_2 = p \sim \mathcal{N}(0, C_{se}([\ell_2, \sigma_2^2]))$ , and  $\eta \sim \mathcal{N}(0, C_{mat, \nu=5/2}([\ell_\eta, \sigma_\eta^2]))$  and (b) a quadratic model:  $y_1 = y_2^2 + \eta$ . The high-order closure is used in the latter experiment. The dashed line represents the “truth” with noisy observations denoted by circles. The solid dark line represents the independent fit and the gray shade the point variance. The blue solid line with blue shade represents the dependent fit.



argue that the positive results obtained in this study extend to other linear and, when using high-order closures, even nonlinear processes. We first generate a synthetic data set in order to validate our covariance models and approach. We then use a real data set from the output of a weather prediction model.

### 3.5.1 Geostrophic Wind

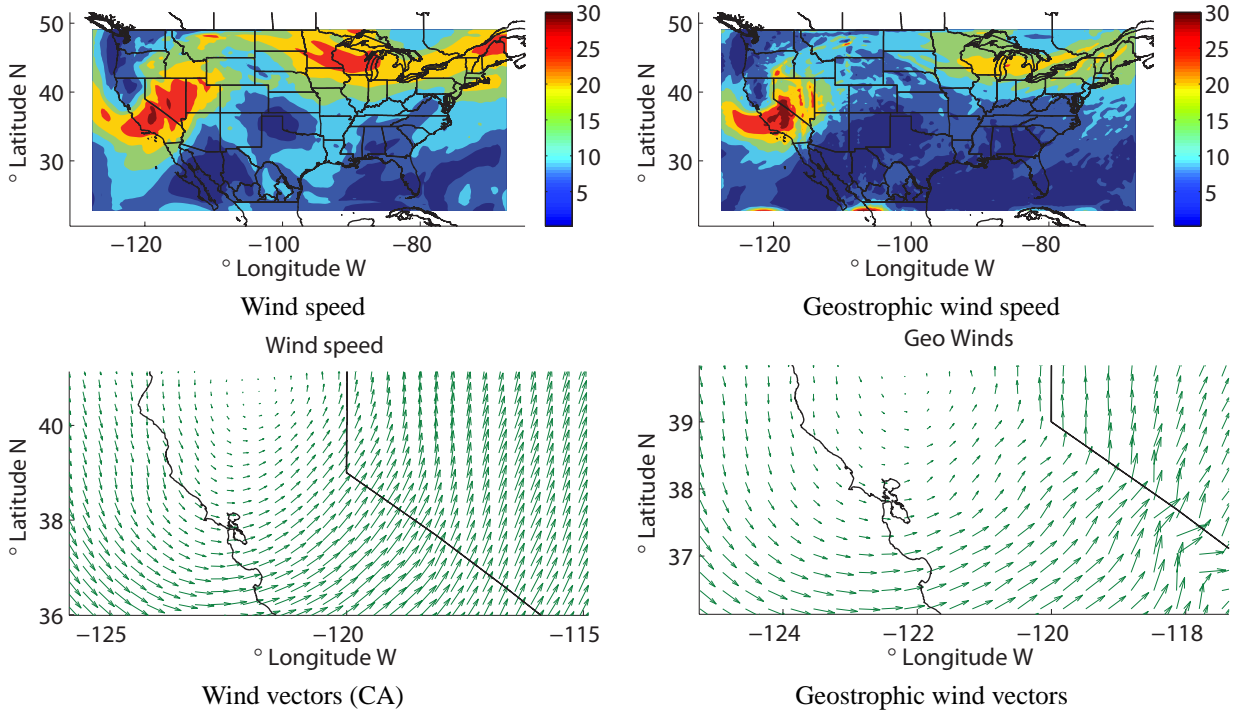
The geostrophic wind is the atmospheric wind field that results from the balance between the Coriolis effect and the pressure gradient force [38]. This is a widely used model for describing the wind fields in the upper troposphere. On a constant pressure surface and on a Cartesian grid, the geostrophic wind follows as

$$u_g = -\alpha_u \frac{\partial \phi_p(x, y)}{\partial y}, \quad v_g = \alpha_v \frac{\partial \phi_p(x, y)}{\partial x}, \quad (24)$$

where  $u_g$  and  $v_g$  are the geostrophic west-east and south-north wind vector components, respectively;  $\phi_p(x, y)$  is the geopotential surface ( $\phi = p/\rho$ , where  $\rho$  is the air density) at a given pressure level  $p$ , and  $\alpha_{u,v}$  is the reciprocal of the Coriolis force. In Fig. 8 we illustrate the wind speed and geostrophic wind approximation in the top panel at a pressure level of 500 mb. In the lower panel we zoom-in over central California to illustrate the relative resemblance between the two fields, but also some discrepancy. We note that in order to obtain the geostrophic wind field in Fig. 8, the scaling  $\alpha_{u,v}$  was manually fitted to a constant for the entire domain, whereas in reality this factor varies in the north-south direction. This approach illustrates two aspects: the differential model is appropriate, and additional forcing is necessary to account for such discrepancies as represented by the apparent vortex disruption.

### 3.5.2 Stochastic Model for Geostrophic Winds

The state vector considered for inference is  $\mathbf{y} = [u_g, v_g, \phi]^T$ . The physics-induced relationship becomes



**FIG. 8:** Wind speed, vector, and geostrophic approximation at a pressure level of 500 mb.



$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} u_g \\ v_g \end{bmatrix} + \Sigma, \quad \Sigma = \mathcal{N}(m_{uv}, \mathbf{K}_{u,v}), \quad \mathbf{K}_{u,v} = \begin{bmatrix} K_{uu} & K_{uv} \\ K_{vu} & K_{vv} \end{bmatrix} \quad (25)$$

$$\mathbf{y}_1 = \begin{bmatrix} u_g \\ v_g \end{bmatrix} = \begin{bmatrix} (-L_y \otimes I_x)\phi \\ (I_y \otimes L_x)\phi \end{bmatrix} = \begin{bmatrix} (-L_y \otimes I_x) \\ (I_y \otimes L_x) \end{bmatrix} (I_2 \otimes \phi) = \begin{bmatrix} L_u \phi \\ L_v \phi \end{bmatrix} = g(\mathbf{y}_2) = g(\phi), \quad (26)$$

where  $L_{x,y}$  is the one-dimensional differential operator given in (24) with respect to the  $x$  (west-east) and  $y$  (south-north) directions, respectively,  $I_{\{x,y\}}$  are identity matrices with dimensions given by the horizontal  $x$  and  $y$  grid points, and  $\otimes$  denotes the Kronecker product.

We propose the following stochastic model to describe the geostrophic process:

$$\phi \sim \mathbf{m}_\phi + \mathcal{M}_{\nu_\phi}(\ell_\phi, \sigma_\phi^2), \quad (27)$$

$$U = L_u \phi + \eta, \quad \mathbf{m}_u = L_u \mathbf{m}_\phi + \mathbf{m}_\eta, \quad \eta \sim \mathcal{M}_{\nu_\eta}(\ell_\eta, \sigma_\eta^2), \quad (28)$$

$$V = L_v \phi + \nu, \quad \mathbf{m}_v = L_v \mathbf{m}_\phi + \mathbf{m}_\nu, \quad \nu \sim \mathcal{M}_{\nu_\nu}(\ell_\nu, \sigma_\nu^2), \quad (29)$$

where  $\mathcal{M}$  is a process generated by using Matérn covariance functions. We consider sparse observations obtained from a synthetically generated data set or the numerical weather model with prescribed additive observational noise described by  $K_{\varepsilon_u \varepsilon_u} = \sigma_u^2 I$ ,  $K_{\varepsilon_v \varepsilon_v} = \sigma_v^2 I$ , and  $K_{\varepsilon_\phi \varepsilon_\phi} = \sigma_\phi^2 I$ , which correspond to wind and geopotential retrievals. A good approximation of these variances may come from the radar and satellite instrumental errors. We will attempt to interpolate the wind and geopotential surfaces using observations with different spatial configurations and densities by using several covariance model structures.

We consider that the predicted quantities (i.e., subscript star) form the “larger” space. With the test-train mappings and notation from previous sections we have  $L_{*u} = H_{u,*u} L_u$ ,  $L_{*v} = H_{v,*v} L_v$ , and  $K_{\phi,\phi} \equiv H_{\phi,*\phi} K_{*\phi,*\phi} H_{\phi,*\phi}^T$ . Then joint distribution of the model proposed for the geostrophic winds takes the following form:

$$\mathcal{M}_3 = \mathbf{K}_{11} = \begin{bmatrix} L_{*u} K_{*\phi,*\phi} L_{*u}^T + K_{\eta\eta} + K_{\varepsilon_u \varepsilon_u} & K_{uv} & L_{*u} K_{*\phi,*\phi} H_{\phi,*\phi}^T \\ K_{vu} & L_{*v} K_{*\phi,*\phi} L_{*v}^T + K_{\nu\nu} + K_{\varepsilon_v \varepsilon_v} & L_{*v} K_{*\phi,*\phi} H_{\phi,*\phi}^T \\ H_{\phi,*\phi} K_{*\phi,*\phi} L_{*u}^T & H_{\phi,*\phi} K_{*\phi,*\phi} L_{*v}^T & K_{\phi,\phi} + K_{\varepsilon_\phi \varepsilon_\phi} \end{bmatrix}, \quad (30)$$

$$\mathbf{K}_{12} = \begin{bmatrix} L_{*u} K_{*\phi,*\phi} L_u^T + K_{\eta,*\eta} & K_{u,*v} & L_{*u} K_{*\phi,*\phi} \\ K_{v,*u} & L_{*v} K_{*\phi,*\phi} L_v^T + K_{\nu,*\nu} & L_{*v} K_{*\phi,*\phi} \\ H_{\phi,*\phi} K_{*\phi,*\phi} L_u^T & H_{\phi,*\phi} K_{*\phi,*\phi} L_v^T & K_{\phi,*\phi} \end{bmatrix}, \quad \mathbf{K}_{21} = \mathbf{K}_{12}^T, \quad (31)$$

$$\mathbf{K}_{22} = \begin{bmatrix} L_u K_{*\phi,*\phi} L_u^T + K_{\eta,\eta} & K_{u,*v} & L_u K_{*\phi,*\phi} \\ K_{v,*u} & L_v K_{*\phi,*\phi} L_v^T + K_{\nu,\nu} & L_v K_{*\phi,*\phi} \\ (L_u K_{*\phi,*\phi})^T & (L_v K_{*\phi,*\phi})^T & K_{\phi,\phi} \end{bmatrix}, \quad (32)$$

where  $K_{uv} = L_{*u} K_{*\phi,*\phi} L_{*v}$  with the rest of the mixed blocks obtained by applying mappings  $H_{u,*u}$  and  $H_{v,*v}$  accordingly. In this example for simplicity we ignore the other terms that would otherwise occur in the expansion of  $K_{uv}$ , such as  $K_{\phi,\eta}$  or  $K_{\eta,\nu}$ . We argue that with sufficient data, one would be able to fit such models; however, this is not the case or scope in our present numerical experiment.

### 3.5.3 Proposed Covariance Models

We distinguish three cases that correspond to (i) an independent fit, (ii) a fit using a latent process, where we consider only wind observations ( $y_1$ ), and (iii) a fit using data from both types of variables. The latent process strategy is similar to the approach discussed in [25] with a slight error in the geostrophic model; however, in our case we consider kernels that couple all observables. The third strategy was used in [17, 26]. In this case the authors provide a particularly specific derivation of the model starting from statistics, whereas our approach starting from the physical constraints arguably carries more generality.

**Independent fit.** In this setting we observe  $U$  and  $V$  as separate processes and try to fit a surface through these observations using Gaussian process regression. To this end we consider the following covariance model:

$$\begin{aligned}\mathcal{M}_2 = \mathbf{K}_{11} &= \begin{bmatrix} K_{*u,*u} + K_{\varepsilon_u \varepsilon_u} & \mathbf{0} \\ \mathbf{0} & K_{*v,*v} + K_{\varepsilon_\phi \varepsilon_\phi} \end{bmatrix}, \\ \mathbf{K}_{12} &= \begin{bmatrix} K_{*u,u} & \mathbf{0} \\ \mathbf{0} & K_{*v,v} \end{bmatrix}, \quad \mathbf{K}_{21} = \mathbf{K}_{12}^T, \\ \mathbf{K}_{22} &= \begin{bmatrix} K_{u,u} & \mathbf{0} \\ \mathbf{0} & K_{v,v} \end{bmatrix}.\end{aligned}$$

We also consider an independent fit of all three quantities. We denote the model by  $\mathcal{M}_4$ . This model is similar to  $\mathcal{M}_2$  but contains an extra block on the diagonal that accounts for the  $\phi$  field.

**Latent process fit.** In this setting we acknowledge that the two components  $U$  and  $V$  are bound together by the geostrophic approximation (24), and therefore  $U$  and  $V$  are components of a joint probability distribution (27)–(29). To this end we obtain the following augmented model:

$$\begin{aligned}\mathcal{M}_1 = \mathbf{K}_{11} &= \begin{bmatrix} L_{*u} K_{*\phi,*\phi} L_{*u}^T + K_{\eta\eta} + K_{\varepsilon_u \varepsilon_u} & L_{*v} K_{*\phi,*\phi} L_{*v}^T + K_{\nu\nu} + K_{\varepsilon_v \varepsilon_v} \\ K_{vu} & \end{bmatrix}, \\ \mathbf{K}_{12} &= \begin{bmatrix} L_{*u} K_{*\phi,*\phi} L_u^T + K_{\eta,*\eta} & K_{u,*v} \\ K_{v,*u} & L_{*v} K_{*\phi,*\phi} L_v^T + K_{\nu,*\nu} \end{bmatrix}, \quad \mathbf{K}_{21} = \mathbf{K}_{12}^T, \\ \mathbf{K}_{22} &= \begin{bmatrix} L_u K_{*\phi,*\phi} L_u^T + K_{*\eta,*\eta} & K_{*u,*v} \\ K_{*v,*u} & L_v K_{*\phi,*\phi} L_v^T + K_{*\nu,*\nu} \end{bmatrix}.\end{aligned}$$

**The fit of the joint process.** This case employs model  $\mathcal{M}_3$ , which corresponds to a regression of the entire data set (30)–(32).

### 3.5.4 Synthetic Example

We first consider a synthetic example with the following hyperparameters:

$$\mathbf{m}_\phi = 0, \quad \nu_\phi = 5/2, \quad \ell_\phi = 10, \quad \sigma_\phi^2 = 30^2, \quad K_{\varepsilon_\phi, \varepsilon_\phi} = 4I, \quad (33)$$

$$\mathbf{m}_\eta = 0, \quad \nu_\eta = 5/2, \quad \ell_\eta = 1, \quad \sigma_\eta^2 = 2, \quad K_{\varepsilon_u, \varepsilon_u} = I, \quad (34)$$

$$\mathbf{m}_\nu = 0, \quad \nu_\nu = 5/2, \quad \ell_\nu = 1, \quad \sigma_\nu^2 = 2, \quad K_{\varepsilon_v, \varepsilon_v} = I, \quad (35)$$

$$\alpha_u = 1.4, \quad \alpha_v = 1.2. \quad (36)$$

The distribution of the complete model  $\mathcal{M}_3$  with these hyperparameters is considered the reference distribution and denoted by  $\mathcal{M}_*$ . We extract two samples from  $\mathcal{M}_*$ : the first is used for fitting the hyperparameters of models  $\mathcal{M}_{\{1\dots 4\}}$ , termed calibration sample, and the second sample, termed validation sample, is used later in a cross-validation experiment. The latter sample is not used in the training phase.

We consider a comprehensive experimental setting that includes two sets of randomly distributed observations in space: a relatively sparse set and a dense one. Furthermore, we consider a secondary experiment in which we consider most of the observations of  $y_1$  to be on the east side (with probability 0.8) and for  $y_2$  on the west side with the same probability. This last setting increases the amount of information transfer between fields  $y_1$  and  $y_2$ , and therefore is expected to increase the discrepancy between models that take into account the physics-induced covariance structure and models that treat the two outputs independently.

In all experiments we show the RMS of the error between predictions and the real value, excluding observed locations. We also show the log-likelihood of the predictions; however, because the covariance models are different and the hyperparameters have different meanings across models, the log-likelihood value has little significance. One exception is when models  $\mathcal{M}_*$  and  $\mathcal{M}_3$  are compared, because they have the same model structure. In Table 1 we show the results from randomly spaced observations. We note that the RMS of the error is larger for the independent fit settings  $\mathcal{M}_2$  and  $\mathcal{M}_4$ . Model  $\mathcal{M}_*$  gives the best fit, which is to be expected since this is the true distribution. The complete model  $\mathcal{M}_3$  performs slightly better than the latent process model  $\mathcal{M}_1$  because of the additional information conveyed by  $y_2$  observations.

We now consider a more extreme case in which the observations are relatively split:  $y_1$  observations are biased toward the eastern side and  $y_2$  observations more toward the western side. The results are shown in Table 2. As expected, we observe a far better performance of the complete model  $\mathcal{M}_3$ . We note a larger discrepancy between  $\mathcal{M}_3$  and the latent process model  $\mathcal{M}_1$  because the latter does not observe  $y_2$  and therefore has relatively few observations on the western side.

### 3.5.5 Real-Data Test Case

We now consider the output of a real numerical weather prediction simulation over North America. We choose a region that is  $54 \times 30$  grid points with a horizontal resolution of 25 km. In Table 3 we show the fit with the four models for the predicted fields from the sample that was used for calibrating the process as well as a validation sample for different noise levels. The validation sample corresponds to the same fields advanced six hours ahead. Because in this case we do not have a reference distribution, we focus more on the RMS of the error between the sample that is used for fitting the hyperparameters (on the left) and the performance on the new sample. A particularly good improvement can be noted in the prediction of  $\phi_*$ .

**TABLE 1:** Predictive marginal log-likelihood values and RMSE (excluding observed locations). This experiment corresponds to the synthetic test case with sparse observations. Observations are random in space for the first two top sets of results.  $\mathcal{M}_*$  [ $y_1y_2$ ] represents the fit with the exact model.  $\mathcal{M}_2$  [ $y_1\_ind$ ] and  $\mathcal{M}_4$  [ $y_1y_2\_ind$ ] are models that treat the fields independently; the former observes only the wind field.  $\mathcal{M}_1$  [ $y_1$ ] and  $\mathcal{M}_3$  [ $y_1y_2$ ] represent the correct model structure and are fitted by using wind for the former and all fields for the latter.

Sparse random observations 0.05% of $U$ , $V$ , and 0.15% of $\phi$								
Model	Calibration sample				Validation sample			
	log-lik	$u_*$	$v_*$	$\phi_*$	log-lik	$u_*$	$v_*$	$\phi_*$
$\mathcal{M}_*$ [ $y_1y_2$ ]	-7798	2.07	1.87	2.47	-7818	2.10	1.94	2.65
$\mathcal{M}_2$ [ $y_1\_ind$ ]	-5865	3.31	2.73	—	-5847	2.69	2.60	—
$\mathcal{M}_4$ [ $y_1y_2\_ind$ ]	-9561	3.34	2.76	3.02	-9527	2.64	2.59	2.91
$\mathcal{M}_1$ [ $y_1$ ]	-5606	2.49	2.41	—	-5662	2.56	2.41	—
$\mathcal{M}_3$ [ $y_1y_2$ ]	-7988	2.33	2.00	2.86	-8026	2.36	2.14	2.96
Dense random observations 0.15% of $U$ , $V$ , and 0.25% of $\phi$								
Model	Calibration sample				Validation sample			
	log-lik	$u_*$	$v_*$	$\phi_*$	log-lik	$u_*$	$v_*$	$\phi_*$
$\mathcal{M}_*$ [ $y_1y_2$ ]	-7041	1.78	1.63	1.98	-7051	1.81	1.66	1.97
$\mathcal{M}_2$ [ $y_1\_ind$ ]	-5220	2.36	2.16	—	-5209	1.99	2.12	—
$\mathcal{M}_4$ [ $y_1y_2\_ind$ ]	-8295	2.34	2.16	2.43	-8242	1.98	2.12	2.39
$\mathcal{M}_1$ [ $y_1$ ]	-4840	1.92	1.77	—	-4880	1.84	1.80	—
$\mathcal{M}_3$ [ $y_1y_2$ ]	-7125	1.85	1.68	2.05	-7166	1.88	1.70	2.03

**TABLE 2:** Predictive marginal log-likelihood values and RMSE (excluding observed locations). Observations are random in space but for  $y_1$  they are predominantly on the east side of the plane with probability 0.8 and on the west side for  $y_2$  with the same probability.  $\mathcal{M}_*$  [y1y2] represents the fit with the exact model.  $\mathcal{M}_2$  [y1\_ind] and  $\mathcal{M}_4$  [y1y2\_ind] are models that treat the fields independently; the former observes only the wind field.  $\mathcal{M}_1$  [y1] and  $\mathcal{M}_3$  [y1y2] represent the correct model structure and are fitted by using wind for the former and all fields for the latter.

<b>Sparse random observations 0.05% of East <math>U</math>, <math>V</math>, and 0.15% of West <math>\phi</math></b>								
	Calibration sample				Validation sample			
Model	log-lik	$u_*$	$v_*$	$\phi_*$	log-lik	$u_*$	$v_*$	$\phi_*$
$\mathcal{M}_*$ [y1y2]	-8146	2.80	2.29	4.89	-8133	2.49	2.38	4.62
$\mathcal{M}_2$ [y1_ind]	-7010	5.56	3.61	—	-6917	3.90	3.84	—
$\mathcal{M}_4$ [y1y2_ind]	-10627	5.58	3.83	9.15	-10478	4.02	4.66	15.18
$\mathcal{M}_1$ [y1]	-5980	4.70	3.18	—	-5973	3.72	3.28	—
$\mathcal{M}_3$ [y1y2]	-8349	2.93	2.44	5.09	-8373	2.88	2.67	6.06

<b>Dense random observations 0.15% of East <math>U</math>, <math>V</math>, and 0.25% of West <math>\phi</math></b>								
	Calibration sample				Validation sample			
Model	log-lik	$u_*$	$v_*$	$\phi_*$	log-lik	$u_*$	$v_*$	$\phi_*$
$\mathcal{M}_*$ [y1y2]	-7772	2.20	1.85	3.23	-7767	2.20	2.00	2.98
$\mathcal{M}_2$ [y1_ind]	-5620	4.08	3.37	—	-5644	2.94	3.09	—
$\mathcal{M}_4$ [y1y2_ind]	-9487	4.08	3.28	5.77	-9399	2.91	2.89	4.45
$\mathcal{M}_1$ [y1]	-5224	2.98	2.57	—	-5277	2.39	2.42	—
$\mathcal{M}_3$ [y1y2]	-7947	2.33	1.97	3.85	-7987	2.32	2.18	3.74

**TABLE 3:** Predictive marginal log-likelihood values and RMSE (excluding observed locations) for the real-data test case with large and small observation noise. The observation density is 0.05% of  $U$ ,  $V$  and 0.15% of  $\phi$ .  $\mathcal{M}_2$  [y1\_ind] and  $\mathcal{M}_4$  [y1y2\_ind] are models that treat the fields independently; the former observes only the wind field.  $\mathcal{M}_1$  [y1] and  $\mathcal{M}_3$  [y1y2] represent the correct model structure and are fitted by using wind for the former and all fields for the latter.

<b>Observation noise: <math>\sigma_\phi^2 = 13^2</math>, <math>\sigma_{\{U,V\}}^2 = 2</math></b>								
	Calibration sample				Validation sample			
Model	log-lik	$u_*$	$v_*$	$\phi_*$	log-lik	$u_*$	$v_*$	$\phi_*$
$\mathcal{M}_2$ [y1_ind]	-5475.19	1.49	1.44	0.00	-5625.26	1.68	1.46	—
$\mathcal{M}_4$ [y1y2_ind]	-15734.91	1.49	1.44	40.75	-16154.14	1.68	1.46	47.75
$\mathcal{M}_1$ [y1]	-4938.89	1.44	1.49	0.00	-5016.47	1.66	1.34	—
$\mathcal{M}_3$ [y1y2]	-16359.38	1.44	1.59	37.37	-17131.86	1.66	1.31	28.72

<b>Observation noise: <math>\sigma_\phi^2 = 8^2</math>, <math>\sigma_{\{U,V\}}^2 = 2</math></b>								
	Calibration sample				Validation sample			
Model	log-lik	$u_*$	$v_*$	$\phi_*$	log-lik	$u_*$	$v_*$	$\phi_*$
$\mathcal{M}_2$ [y1_ind]	-5475	1.49	1.44	—	-5625	1.68	1.46	—
$\mathcal{M}_4$ [y1y2_ind]	-14200	1.46	1.39	33.55	-14363	1.65	1.40	44.74
$\mathcal{M}_1$ [y1]	-5289	1.43	1.38	—	-5384	1.62	1.28	—
$\mathcal{M}_3$ [y1y2]	-13129	1.43	1.38	22.02	-13713	1.63	1.28	29.48

Although the results for the large observational noise do not mirror precisely the ones obtained in the synthetic-data case, we note that reducing the noise level (results in the lower part of the table) lead to the same conclusions that were drawn in the previous section. We adopted several levels of simplifications in the covariance models and calibration strategies employed in the real data case. For instance, the Coriolis force varies in the north-south direction; however,  $\alpha_u$  and  $\alpha_v$  are kept constant across the entire domain. Moreover, all the processes are Matérn with a fixed smoothness level of  $\nu = 5/2$ . We expect more accurate results to be obtained by adding more flexibility to the inference process; however, this is the scope of a different study.

To give a slightly more qualitative representation of the results, we illustrate in Fig. 9 the geopotential error surface corresponding to predictions made on the validation sample in Table 3 using the independent fit  $\mathcal{M}_4$  [y1y2\_ind] and  $\mathcal{M}_3$  [y1y2] models with  $\sigma_\phi^2 = 8^2$ ,  $\sigma_{\{U,V\}}^2 = 2$ . A general and significant reduction in the prediction error can be noticed when using the physics-based model with a significant level on the western boundary.

### 3.5.6 Validation of Covariance Models

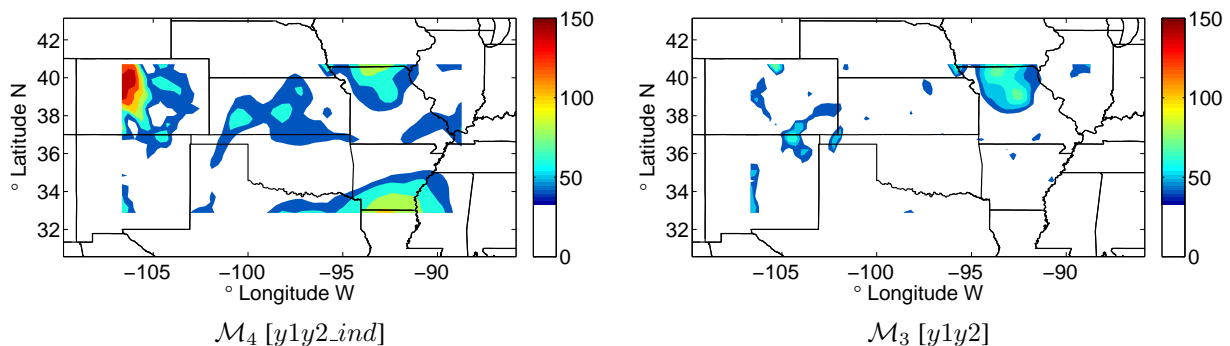
We propose two approaches to validate the calibrated models. The first approach is based on a cross-validation strategy, where we draw a second sample from the same distribution or process and perform the regression by using the models calibrated on the initial sample. In the synthetic data case we use a different seed, and in the real-data experiment we use a different time snapshot of the geopotential and wind fields.

In a second validation approach we take advantage of the fact that in the synthetic data experiment we already know the true distribution, and therefore we can construct the true Gaussian process according to (30)–(32) and with hyperparameters given by (33)–(36). We then compute the Kullback-Leibler (KL) divergence between the distributions resulting from the different models calibrated with the data and the true distribution. This gives us a “measure” of the distance between the experimentally calibrated distributions and the true one. The KL divergence between two probability densities that are normally distributed,  $p = \mathcal{N}(\mu_p, \Sigma_p)$  and  $q = \mathcal{N}(\mu_q, \Sigma_q)$ , is given by

$$\begin{aligned} \mathcal{D}_{\text{KL}}(p||q) &= \int p(x) \log \frac{p(x)}{q(x)} dx \\ &= \frac{1}{2} \left( \text{tr}(\Sigma_p \Sigma_q^{-1}) + (\mu_q - \mu_p)^T \Sigma_q^{-1} (\mu_q - \mu_p) - \ln \left( \frac{|\Sigma_p|}{|\Sigma_q|} \right) - N \right), \end{aligned} \quad (37)$$

where  $N$  is the dimension of the problem.

**First Approach: One-Way Cross Validation.** The results for the first approach have already been discussed to a certain extent. In Tables 1–3 we illustrate (in the right set of columns) the validation sample results, which correspond to a new sample from the same distribution in the synthetic-data case or to a different time snapshot in the real-data



**FIG. 9:** Geopotential error surface [ $m^2 s^{-2}$ ] corresponding to predictions made on the validation sample in the real-data experiment (see Table 3 for the independent fit  $\mathcal{M}_4$  [y1y2\_ind] and  $\mathcal{M}_3$  [y1y2] in the case where  $\sigma_\phi^2 = 8^2$ ,  $\sigma_{\{U,V\}}^2 = 2$ ).

case. This validation sample was not used in fitting the hyperparameters. In all cases when using the appropriate covariance structure  $\mathcal{M}_1$  and  $\mathcal{M}_3$  maintain an advantage over independent fit, which indicates their robustness. Note the columns that correspond to the validation sample in Table 2 and the fit of  $\phi$  in Table 3. Also, there is no significant change in the RMS of the error between the calibration sample and the predicted one, which may discard the possibility of overfitting the models.

**Second Approach: KL Divergence Experiment.** Now we compute the KL divergence (37) for the synthetic data case, where  $q$  takes the place of the known distribution with hyperparameters (33)–(36) and  $p$  takes the place of distributions generated by  $\mathcal{M}_{\{1,\dots,4\}}$  and inferred parameters. In Table 4 we show the KL divergence between the four models  $\mathcal{M}_{\{1,\dots,4\}}$  and the true distribution denoted by  $\mathcal{M}_*$ . By  $\mathcal{D}_{\text{KL}}(\mathcal{M}_k||\mathcal{M}_*)$ ,  $k = 1, \dots, 4$ , we indicate the KL divergence of each model with respect to the true distribution, and by  $\mathcal{D}_{\text{KL}}(\int \mathcal{M}_k d\phi||\int \mathcal{M}_* d\phi)$  we indicate the marginal with respect to  $\phi$ . The latter is used to compare models that include  $\phi$  with the ones that do not; for instance this allows us to compare directly  $\mathcal{M}_1$  and  $\mathcal{M}_3$ . These results correspond to the models presented in Table 1; similar results are obtained for the models obtained in Table 2. We note the relative closeness between  $\mathcal{M}_1$  and  $\mathcal{M}_3$ , a fact expected from the forecast fit result.

#### 4. DISCUSSION

The covariance structure has a large impact on the uncertainty quantification and forecast efficiency. Auto-covariance and cross-covariance models are needed to represent joint distributions of random fields that may be generated from physical fields that have different meanings or interpretations but are constrained by physical laws. In particular, having a consistent covariance structure is known to be important for prediction when performing inferences on multidimensional process with partially known relationships among different variables.

In this study we propose covariance models that are consistent with the underlying physical process that generated the data. The covariance model describes how the outputs covary and may have nontrivial forms when relating different physical quantities. This study is geared toward covariance models that describe data obtained from processes that obey at least a partially known underlying physical process. With such a suitable covariance structure, one can make predictions using Gaussian process regression strategies or employ them in other circumstances to describe uncertainties in models, modeling, and data sets.

We develop analytical covariance functions that are consistent with several physical processes. In particular, we focus on modeling the geostrophic wind in the atmosphere, and to that end we employ a differential process that corresponds to the known physical constraint. We consider Gaussian process regression experiments with a covariance model that has the correct physically consistent structure, which demonstrates significant improvements in the forecast efficiency. This strategy is validated on various synthetic and real data sets. The analytic covariance functions are validated by comparing results obtained with the models introduced in this study and covariance structures obtained through sampling strategies.

We introduce new nonstationary covariance models that are generated directly through the physical process. For instance, we use a differential model on a nonuniform grid to generate nonstationary covariance kernels. These mod-

**TABLE 4:** Kullback-Leibler divergence between the four models  $\mathcal{M}_{\{1,\dots,4\}}$  and the true distribution  $\mathcal{M}_*$ . By  $\mathcal{D}_{\text{KL}}(\mathcal{M}_k||\mathcal{M}_*)$ ,  $k = 1, \dots, 4$  we indicate the KL divergence of each model with respect to the true distribution and by  $\mathcal{D}_{\text{KL}}(\int \mathcal{M}_k d\phi||\int \mathcal{M}_* d\phi)$  we indicate the marginal with respect to  $\phi$ . The latter is used to compare models that include  $\phi$  with the ones that do not. The results are based on the  $K_{22}$  block.

Model	Sparse observations		Dense observation	
	$\mathcal{D}_{\text{KL}}(\mathcal{M}_k  \mathcal{M}_*)$	$\mathcal{D}_{\text{KL}}(\int \mathcal{M}_k d\phi  \int \mathcal{M}_* d\phi)$	$\mathcal{D}_{\text{KL}}(\mathcal{M}_k  \mathcal{M}_*)$	$\mathcal{D}_{\text{KL}}(\int \mathcal{M}_k d\phi  \int \mathcal{M}_* d\phi)$
$\mathcal{M}_2$ [y1_ind]	—	4662	—	4197
$\mathcal{M}_4$ [y1y2_ind]	50564	4320	34541	4016
$\mathcal{M}_1$ [y1]	—	746	—	127
$\mathcal{M}_3$ [y1y2]	3217	798	778	215

els have properties that are appropriate for processes that take place on adaptive grids or have various degrees of anisotropy.

We have augmented our analysis to include covariance models that are able to effectively describe nonlinear processes by including high-order correction terms, which can be regarded as high-order moment closure terms. We have demonstrated that such a strategy, albeit not an optimal one for our particular experiment, can be very important for nonlinear models by comparing the fine approximation of the covariance structure resulting from a nonlinear process with low- and high-order closure assumptions. The latter proves to be significantly more accurate.

Gaussian processes are known to be practical as long as one can perform the Cholesky decomposition of the covariance matrix, but for very-large-scale data sets this approach may become a problem limiting their applicability. In this study we do not fully address the computational aspects that are involved in the Gaussian process regression; however, recent results [49] demonstrate that Gaussian process analysis can be carried out in a matrix-free fashion in a way that scales very well and, therefore, can be applied to large-scale problems.

## ACKNOWLEDGMENTS

This work was supported by the US Department of Energy through contract no. DE-AC02-06CH11357. We thank Michael Stein for comments on multiple versions of this manuscript.

## REFERENCES

1. Williams, C., Prediction with Gaussian processes: From linear regression to linear prediction and beyond, in *Learning and Inference in Graphical Models*, pp. 599–621, Kluwer, Dordrecht, 1998.
2. MacKay, D., Introduction to Gaussian processes, Tech. Rep. pp. 133–165, ATO ASI series F: Computer and System Sciences, 1998.
3. Rasmussen, C. and Williams, C., *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press, Cambridge, MA, 2005.
4. Higdon, D., *Statistical Methods for Spatio-Temporal Systems*, chapter VI, A primer on space-time modeling from a Bayesian perspective, pp. 217–279, Chapman & Hall, London, 2007.
5. Stein, M., *Interpolation of Spatial Data: Some Theory for Kriging*, Springer Verlag, Berlin, 1999.
6. Kocijan, J., Murray-Smith, R., Rasmussen, C., and Girard, A., *Gaussian process model based predictive control*, in American Control Conf., Boston, 2004.
7. Chiles, J. and Delfiner, P., *Geostatistics: Modeling Spatial Uncertainty*, Wiley-Interscience, New York, 1999.
8. Boyle, P. and Frean, M., Dependent Gaussian processes, in *Advances in Neural Information Processing Systems 17: Proc. of the 2004 Conf.*, pp. 217–224, The MIT Press, Cambridge, MA, 2005.
9. Girard, A., Rasmussen, C., and Murray-Smith, R., Multiple-step ahead prediction for nonlinear dynamic systems—A Gaussian process treatment with propagation of the uncertainty, Tech. Rep. TR-2002-119, University of Glasgow, 2002.
10. Girard, A., Rasmussen, C., Candela, J., and Murray-Smith, R., Gaussian process priors with uncertain inputs-application to multiple-step ahead time series forecasting, in *Advances in Neural Information Processing Systems*, pp. 545–552, The MIT Press, Cambridge, MA, 2003.
11. Candela, J., Girard, A., and Rasmussen, C., Prediction at an uncertain input for Gaussian processes and relevance vector machines application to multiple-step ahead time-series forecasting, Tech. Rep., Technical University of Denmark, 2003.
12. Osborne, M., Gaussian processes for prediction, Tech. Rep. PARG-07-01, University of Oxford, 2007.
13. Wang, J., Fleet, D., and Hertzmann, A., Gaussian process dynamical models for human motion, *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):283–298, 2007.
14. Cressie, N. and Wikle, C., *Statistics for Spatio-Temporal Data*, Wiley, New York, 2011.
15. Boyle, P. and Frean, M., Multiple output Gaussian process regression, Tech. Rep., Victoria University of Wellington, 2005.
16. Stein, M., The loss of efficiency in kriging prediction caused by misspecifications of the covariance structure, *Geostatistics*, 1:273–282, 1989.

17. Chauvet, P., Pailleux, J., and Chiles, J., Analyse objective des champs météorologiques par cokrigage, *La Météorologie, Sci. Tech.*, 6:37–54, 1976.
18. Cohn, S., Dynamics of short-term univariate forecast error covariances, *Mon. Weather Rev.*, 121(11):3123–3149, 1993.
19. Gaspari, G. and Cohn, S., Construction of correlation functions in two and three dimensions, *Q. J. R. Meteorol. Soc.*, 125:723–757, 1999.
20. Gaspari, G., Cohn, S., Guo, J., and Pawson, S., Construction and application of covariance functions with variable length-fields, *Q. J. R. Meteorol. Soc.*, 132(619):1815–1838, 2006.
21. Gelman, A., Carlin, J., Stern, H., and Rubin, D., *Bayesian Data Analysis*, Chapman & Hall, London, 2nd edition, 2003.
22. Berliner, L., Hierarchical Bayesian time-series models, in: Hanson, K. and Silver, R. (Eds.), *Maximum Entropy and Bayesian Methods*, Vol. 79, pp. 15–22, Kluwer Academic Publishers, Dordrecht, 1996.
23. Wikle, C., Berliner, L., and Cressie, N., Hierarchical Bayesian space-time models, *Environ. Ecol. Stat.*, 5(2):117–154, 1998.
24. Berliner, L., Royle, J., Wikle, C., and Milliff, R., Bayesian methods in the atmospheric sciences, in *Bayesian Statistics 6: Proc. of the Sixth Valencia International Meeting, June 6–10, 1998*, pp. 83–100, Oxford University Press, New York, 1999.
25. Royle, J., Berliner, L., Wikle, C., and Milliff, R., A hierarchical spatial model for constructing wind fields from scatterometer data in the Labrador Sea, *Case Studies Bayesian Stat.*, 4:367–382, 1999.
26. Chiles, J. and de Paris, S., How to adapt kriging to non-classical problems: Three case studies, 24:69–89, 1976.
27. Berliner, L., Physical-statistical modeling in geophysics, *J. Geophys. Res.-Atmos.*, 108(D24):8776, 2003.
28. Berliner, L., Milliff, R., and Wikle, C., Bayesian hierarchical modeling of air-sea interaction, *J. Geophys. Res.*, 108:3104–3120, 2003.
29. Wikle, C., Milliff, R., Nychka, D., and Berliner, L., Spatiotemporal hierarchical Bayesian modeling tropical ocean surface winds, *J. Am. Stat. Assoc.*, 96(454):382–397, 2001.
30. Clark, J. and Gelfand, A., *Hierarchical Modelling for the Environmental Sciences: Statistical Methods and Applications*, Oxford University Press, New York, 2006.
31. Lee, H., Higdon, D., Calder, C., and Holloman, C., Efficient models for correlated data via convolutions of intrinsic processes, *Stat. Model.*, 5(1):53–74, 2005.
32. Campbell, E., Statistical modeling in nonlinear systems, *J. Climate*, 18(16):3388–3399, 2005.
33. Constantinescu, E., Chai, T., Sandu, A., and Carmichael, G., Autoregressive models of background errors for chemical data assimilation, *J. Geophys. Res.*, 112:D12309, 2007.
34. Wan, F., Linear partial differential equations with random forcing, *Stud. Appl. Math.*, 51(2):163–178, 1972.
35. Apanasovich, T. and Genton, M., Cross-covariance functions for multivariate random fields based on latent dimensions, *Biometrika*, 97(1):15, 2010.
36. Gneiting, T., Kleiber, W., and Schlather, M., Matérn cross-covariance functions for multivariate random fields, *J. Am. Stat. Assoc.*, 105(491):1167–1177, 2010.
37. Zhang, Z., Beletsky, D., Schwab, D., and Stein, M., Assimilation of current measurements into a circulation model of Lake Michigan, *Water Resour. Res.*, 43(11):W11407, 2007.
38. Holton, J., *An Introduction to Dynamic Meteorology*, Academic Press, London, 2004.
39. Lockwood, B. and Anitescu, M., Gradient-enhanced universal kriging for uncertainty propagation in nuclear engineering, *Trans. Am. Nucl. Soc.*, In press, 2011.
40. Boyd, S. and Vandenberghe, L., *Convex Optimization*, Cambridge University Press, Cambridge, 2004.
41. Horn, R. and Johnson, C., *Matrix Analysis*, Cambridge University Press, Cambridge, 2005.
42. Oehlert, G., A note on the Delta method, *Am. Stat.*, 46(1):27–29, 1992.
43. Casella, G., Berger, R., and Berger, R., *Statistical Inference*, Duxbury, Pacific Grove, CA, 2002.
44. Boyce, W., Random eigenvalue problems, *Probab. Methods Appl. Math.*, 1:1–73, 1968.
45. Crandall, S., Non-Gaussian closure techniques for stationary random vibration, *Int. J. Nonlinear Mech.*, 20(1):1–8, 1985.
46. Crandall, S., Non-Gaussian closure for random vibration of non-linear oscillators, *Int. J. Nonlinear Mech.*, 15(4-5):303–313, 1980.



47. Ibrahim, R., Soundararajan, A., and Heo, H., Stochastic response of nonlinear dynamic systems based on a non-Gaussian closure, *J. Appl. Mech.*, 52:965–970, 1985.
48. Isserlis, L., On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables, *Biometrika*, 12(1-2):134, 1918.
49. Anitescu, M., Chen, J., and Wang, L., A matrix-free approach for solving the Gaussian process maximum likelihood problem, Argonne National Laboratory, Preprint ANL/MCS–P1857–0311, 2011.