RESEARCH ARTICLE

WILEY

The scope of the Kalman filter for spatio-temporal applications in environmental science

Jonathan Rougier[1] | Aoibheann Brady[2,3] | Jonathan Bamber[2] | Stephen Chuter[2] | Sam Royston[2] | Bramha Dutt Vishwakarma[2,4,5] | Richard Westaway[2] | Yann Ziegler[2]

[1]School of Mathematics, University of Bristol, Bristol, UK

[2]School of Geographical Sciences, University of Bristol, Bristol, UK

[3]Cervest, London, UK

[4]Interdisciplinary Centre for Water Research, Indian Institute of Science, Bangalore, India

[5]Centre for Earth Science, Indian Institute of Science, Bangalore, India

Correspondence
Jonathan Rougier, School of Mathematics, University of Bristol, Bristol, UK.
Email: j.c.rougier@bristol.ac.uk

Abstract

The Kalman filter is a workhorse of dynamical modeling. But there are challenges when using the Kalman filter in environmental science: the complexity of environmental processes, the complicated and irregular nature of many environmental datasets, and the scale of environmental datasets, which may comprise many thousands of observations per time-step. We show how these challenges can be met within the Kalman filter, identifying some situations which are relatively easy to handle, such as datasets which are high-resolution in time, and some which are hard, like areal observations on small contiguous polygons. Overall, we conclude that many applications in environmental science are within the scope of the Kalman filter, or its generalizations.

KEYWORDS

areal observations, basis expansion, local linear trend, sequential updating, truncation error, upscaling

1 | INTRODUCTION

The Kalman filter (Kalman, 1960) has been a workhorse of dynamical modeling for half a century. Despite enormous changes in computing power and architecture, it is still a core tool in statistics, including spatio-temporal modeling (Cressie & Wikle, 2011), and machine learning (Murphy, 2012). The "vanilla" Kalman filter has a linear state equation and a linear observation equation. It is a platform for many generalizations. For example, the Ensemble Kalman filter (EnKF, Evensen, 2007; Katzfuss et al., 2016) and the Unscented Kalman filter (Julier & Uhlmann, 1997) accommodate a nonlinear state equation. There are various approaches to accommodate a nonlinear observation equation, including the EnKF and dynamical Generalized Linear Models (Katzfuss et al., 2020).

This article explores the scope of the vanilla Kalman filter for applications in environmental science. We examine three fundamental challenges. First, environmental processes are complex, and may require rich highly-parameterized models in order to capture dynamical behavior over a spatially diverse domain. Second, environmental datasets are complicated, and often fail to conform to the simple temporal regularity of the Kalman filter, which proceeds in a sequence of equally-spaced time-steps. Third, environmental datasets can be "massive", by which we mean that they may comprise thousands or tens of thousands of observations per time-step.

The outline of the article is as follows. Section 2 presents a simplified version of an application we have been working on for several years, to provide illustrations of the various challenges. Sections 3 and 4 present our vanilla Kalman

filter, which captures a wide variety of model structures suitable for environmental science. Section 3 states the model and the computational cost of inference; Section 4 reviews some aspects of the model that are relevant for environmental science. In particular, this section introduces the "sequential update" (SU) condition which is necessary for massive datasets. Sections 5 and 6 are about "upscaling" datasets which are high-resolution in time and in space. The first of these does not present too many difficulties, but the second can violate the SU condition, and so needs to be carefully implemented for massive datasets. Section 7 is about areal observations, where violating the SU condition is more difficult to avoid. Section 8 generalizes the model of Section 3 to allow for local trends not captured by covariates, and also considers differencing as a "quick-and-dirty" approximation. Section 9 is a brief conclusion, including a summary of the argument in the paper.

2 | APPLICATION

Here is a skeleton application, which is already rich in complications, and which reflects the type of inference involved in a project aimed at resolving and attributing contributions to sea level rise at the global scale (the GlobalMass project, https://www.globalmass.eu/). The latent process of interest is North American land elevation relative to the WGS-84 ellipsoid (see, e.g., Kaplan & Hegarty, 2017).

There are two sources of data. First, global positioning system (GPS) stations which report their time, longitude, latitude, and elevation (see, e.g., Blewitt et al., 2018). These are irregularly located in space and time, although when a GPS station is operating at a site it tends to run for several years, and report roughly daily. It is reasonable to assume that the observation errors at well-separated stations are independent, although observations at nearby stations might be correlated by local processes, including those connected with urbanization.

Second, data from the GRACE satellites (see, e.g., Watkins et al., 2015). GRACE denotes "Gravity Recovery and Climate Experiment". This pair of satellites measures gravity anomalies with a typical spatial resolution of about 300 km, and these can be converted to vertical land motion (VLM) by making some assumptions about the density of the lithosphere and mantle. After processing, the GRACE dataset takes the form of monthly changes in monthly mean elevations over specified spatial polygons, and is available for a set of polygons which tile North America each with typical width of a few 100 km. Since the GRACE observations are extensively processed from their raw state, it is questionable to assume that the observation errors for GRACE are independent in space or in time on small polygons, although this dependence will be diminished by aggregating to annual differences, and to larger polygons.

Much more could be said about both of these datasets, which would be important in practice, but not germane for this article; for example, for GPS, see He et al. (2020), while for GRACE, see Chao (2016) and Vishwakarma et al. (2022). We will refer to both datasets as "observations", recognizing that in environmental science, preprocessing of raw measurements into quantities that might have been observed is standard, and there is little benefit in distinguishing between "true" observations and derived or analogue observations.

3 | THE KALMAN FILTER

This section states our statistical model, a "vanilla" Kalman filter, and outlines how the cost of inference scales with the size of the dataset, and the complexity of the model. Section 4 reflects on the use of the model in environmental science, and Section 8 gives an important generalization. For familiarity, the notation is similar to Cressie and Wikle (2011, ch. 7).

3.1 | The model

There is a spatio-temporal latent process of interest, denoted Y, where the value of Y at location $s \in \mathcal{D}$ and time $t \in \{0, \ldots, T\}$ is denoted $Y_t(s)$. That is, Y is treated as discrete in time, for modeling purposes; nevertheless, the underlying Y may be continuous in time (see Section 5). For simplicity, treat Y as a scalar process: for the application in Section 2, Y is elevation. In the sections that follow we will assume one latent process and one type of dataset, but in practice the model allows multiple latent processes and multiple types of dataset, which can be stacked together within the same model structure.

First, assume that the latent process Y can be represented as a basis expansion,

$$Y_t(s) - \bar{y}_t(s) = \sum_{j=1}^{d} \phi_j(s) \cdot \alpha_{tj} + v_t(s) =: \boldsymbol{\phi}(s)^T \boldsymbol{\alpha}_t + v_t(s), \tag{1}$$

where $\bar{y}$ is a mean function for Y, $\boldsymbol{\phi}$ is a specified finite set of time-invariant basis functions with coefficients $\boldsymbol{\alpha}_t$, and v is a spatio-temporal Gaussian process which represents truncation error. Thus Y is linearly related to the coefficients $\boldsymbol{\alpha}$, which will be discussed further in Section 4. There is no presumption that $\boldsymbol{\alpha}_t$ is a small set of coefficients; it might have thousands of components, in an application like that in Section 2.

In (1), the mean function $\bar{y}$ may be specified, or it may itself have uncertain coefficients; for example,

$$\bar{y}_t(s) = \beta_0 + \sum_{j=1}^{k} x_j(s, t) \cdot \beta_j, \tag{2}$$

where $\boldsymbol{x}$ represents specified covariates, and the $\boldsymbol{\beta}$ coefficients are treated as time-invariant. Covariates can be crucial in practice; for example, they can account for large-scale effects and discontinuities, and thus allow the basis functions to be smooth and localized (Bolin et al., 2019). Alternatively, the $\boldsymbol{\beta}$ coefficients could vary through time, like $\boldsymbol{\alpha}_t$; or some of the $\boldsymbol{\alpha}_t$ might be time-invariant. These generalizations do not create any complications and will be ignored in what follows.

Let $s_1, \ldots, s_m$ be any set of locations in $\mathcal{D}$. Define

$$\boldsymbol{Y}_t := \begin{bmatrix} Y_t(s_1) \\ \vdots \\ Y_t(s_m) \end{bmatrix}, \quad \bar{\boldsymbol{y}}_t := \begin{bmatrix} \bar{y}_t(s_1) \\ \vdots \\ \bar{y}_t(s_m) \end{bmatrix}, \quad \boldsymbol{v}_t := \begin{bmatrix} v_t(s_1) \\ \vdots \\ v_t(s_m) \end{bmatrix} \tag{3a}$$

and

$$\Phi_t := \begin{bmatrix} \phi_1(s_1) & \ldots & \phi_d(s_1) \\ \vdots & \ddots & \vdots \\ \phi_1(s_m) & \ldots & \phi_d(s_m) \end{bmatrix}. \tag{3b}$$

Then (1) implies that

$$\boldsymbol{Y}_t = \bar{\boldsymbol{y}}_t + \Phi_t \boldsymbol{\alpha}_t + \boldsymbol{v}_t. \tag{3c}$$

Assume that the basis expansion in (1) is sufficiently rich that a simple statistical model suffices for the truncation error v, independent of $\boldsymbol{\alpha}$, with expectation 0 and

$$\text{Cov}\left(v_t(s), v_{t'}(s')\right) = \begin{cases} \kappa^2 \cdot C(s, s') & t = t', \\ 0 & t \neq t', \end{cases} \tag{4}$$

where κ is a parameter and C is a spatial correlation function. This assumption implies that

$$\text{Var}(\boldsymbol{v}_t) = \kappa^2 C_t, \tag{5}$$

where C_t is the $m \times m$ *Gram matrix* for C at locations $s_1, \ldots, s_m$; that is,

$$(C_t)_{ij} := C(s_i, s_j), \quad i, j = 1, \ldots, m.$$

The truncation error v is a source of difficulty, and will reoccur several times below.

Dynamics for Y are induced by a time-invariant Gaussian dynamical model for $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}_t = M\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t \quad \boldsymbol{\eta}_t \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{0}, Q)$$

$$\boldsymbol{\alpha}_0 \overset{\text{ind}}{\sim} \mathcal{N}(m_0, S_0), \tag{6}$$

where the matrices M and Q are parameters, or parameterized in some fashion. Assume that the spectral radius of M (the largest modulus of the eigenvalues of M) is less than 1, which implies that both $\mathrm{E}(\boldsymbol{\alpha}_t)$ and $\mathrm{Var}(\boldsymbol{\alpha}_t)$ have stationary values (Wendland, 2018, ch. 4). These are easily found to be

$$m_0 = \mathbf{0}, \quad S_0^V = (I - M \otimes M)^{-1} Q^V, \tag{7}$$

where S_0^V is the vectorization of S_0 by column, similarly for Q^V, and $\otimes$ is the Kronecker product (see, e.g., Mardia et al., 1979, appendix A). Setting m_0 and S_0 in this way eliminates the need to specify or estimate a mean and variance for $\boldsymbol{\alpha}_0$, and make the prior process for $\boldsymbol{\alpha}_t$ stationary. In this model, with stationary $\boldsymbol{\alpha}_t$, Y_t will tend to $\bar{y}_t$, in the absence of shocks.

Let

$$\boldsymbol{Z}_t := \begin{bmatrix} Z_{t1} \\ \vdots \\ Z_{tn} \end{bmatrix}, \quad \boldsymbol{\epsilon}_t := \begin{bmatrix} \epsilon_{t1} \\ \vdots \\ \epsilon_{tn} \end{bmatrix} \tag{8}$$

denote a set of n observables at time-step t, with observation errors $\boldsymbol{\epsilon}_t$. We write "observables" for random variables, and "observations" for the measured values of those random variables, where this distinction is meaningful. Assume there exists a specified set of locations $s_1, \ldots, s_m$, a specified $n \times m$ incidence matrix H_t, and a specified variance matrix Σ_t, for which

$$\boldsymbol{Z}_t = H_t \boldsymbol{Y}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma_t). \tag{9}$$

We term (9) the *direct observation equation*. In the simplest case, $Z_{ti} = Y_t(s_i) + \epsilon_{ti}$, and $H_t = I$, but more complicated cases are necessary, as discussed below. Along with H_t and Σ_t, both $s_1, \ldots, s_m$ and n can vary in time, but this has been suppressed in the notation to reduce clutter.

The observation error variance Σ_t is usually treated as diagonal. In fact there are practices which suggest that Σ_t ought not to be diagonal, such as data processing and smoothing, but the effect of these practices is seldom quantified and passed on to users of the observations. Assume that Σ_t is diagonal, for simplicity.

Expanding (9) out using (3),

$$\boldsymbol{Z}_t = H_t(\bar{\boldsymbol{y}}_t + \Phi_t \boldsymbol{\alpha}_t + \boldsymbol{v}_t) + \boldsymbol{\epsilon}_t. \tag{10}$$

Collecting the stochastic terms together gives the final form of the observation equation,

$$\boldsymbol{Z}_t = H_t(\bar{\boldsymbol{y}}_t + \Phi_t \boldsymbol{\alpha}_t) + \boldsymbol{\gamma}_t, \quad \boldsymbol{\gamma}_t \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{0}, R_t) \tag{11a}$$

with

$$R_t := \kappa^2 H_t C_t H_t^T + \Sigma_t, \tag{11b}$$

where C_t is the Gram matrix of the correlation function C at the locations $s_1, \ldots, s_m$, as above. We term (11) the *indirect observation equation*. It is helpful to make this distinction between the natural relationship connecting the observables and the latent process (direct), and the model-based relationship connecting the observables and the state vector (indirect).

This completes the dynamical spatio-temporal model (DSTM) for the latent process Y and the observables

$$\boldsymbol{Z}_{1:T} := (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T),$$

which comprises the basis representation (1), the state equation (6), and the direct observation equation (9), although in practice the last of these is replaced by the indirect observation equation (11). The mean function $\bar{y}$ and the basis functions $\boldsymbol{\phi}$ are specified, as are the incidence matrices $H_1, \ldots, H_T$ and the (diagonal) observation error variances $\Sigma_1, \ldots, \Sigma_T$, along with the locations necessary to instantiate them. The parameters of the DSTM are M and Q from the state equation, and κ and C, representing truncation error in the basis expansion.

Various generalizations of this model are possible. In the state equation it is not necessary that M is time-invariant, and if it is time-invariant, it is not necessary that M is stationary: these assumptions are made for simplicity, and to reduce the number of parameters that need to be estimated. On the other hand, it is crucial that the truncation error in the observation equation is localized in time, because it is fundamental to the Kalman filter that time-dependence passes through the mean function and the coefficients. Having said that, the hope is that the basis expansion for the latent process is sufficiently rich that the detailed structure of the truncation error does not matter for the inference.

3.2 | Inference

Suppose that inference comprises estimating point values for the parameters $\theta := \{M, Q, \kappa, C\}$ by maximum likelihood, denoted $\hat{\theta}$, and conditional distributions for $Y_0, \ldots, Y_T$ given observations $\boldsymbol{z}_{1:T}^{\text{obs}}$, using the plug-in $\theta = \hat{\theta}$. This can be carried out using the equations of the Kalman filter; see Murphy (2012, sec. 18.3) or Cressie and Wikle (2011, sec. 8.2). The Kalman filter is used here to provide a benchmark for computational cost, and to highlight those features of the observables which compromise computational efficiency.

The Kalman filter provides the sequence of filtering distributions

$$p_\theta\big(\boldsymbol{\alpha}_t \mid \boldsymbol{z}_{1:t}^{\text{obs}}\big), \quad t = 1, \ldots, T,$$

which, in the linear Gaussian DSTM of Section 3.1, are Gaussian distributions represented by expectation vectors and variance matrices. The baseline cost of filtering is $O(Tn^3)$, where n is the number of observations in each step (assumed to be the same each time-step, to reduce clutter). The Kalman filter also provides the log-likelihood value

$$\log L(\theta) := \log p_\theta\big(\boldsymbol{z}_{1:T}^{\text{obs}}\big)$$

at no extra cost.

After filtering, the Kalman smoother provides the sequence of smoothing distributions

$$p_\theta\big(\boldsymbol{\alpha}_t \mid \boldsymbol{z}_{1:T}^{\text{obs}}\big), \quad t = T, \ldots, 0,$$

each of which is also Gaussian. The cost of smoothing is $O(Td^3)$, where d is the length of the state vector $\boldsymbol{\alpha}_t$, a measure of model complexity. Applying the basis expansion, (1), the Kalman smoother also provides

$$p_\theta\big(\boldsymbol{Y}_t \mid \boldsymbol{z}_{1:T}^{\text{obs}}\big), \quad t = 0, \ldots, T,$$

where $\boldsymbol{Y}_t$ is the latent process at time t at any finite set of sites, also Gaussian. The truncation error ν is not updated, but its correlation structure contributes to the covariance of $\boldsymbol{Y}_t \mid \boldsymbol{z}_{1:T}^{\text{obs}}$. In principle, ν could be also be updated, which would decrease the smoothing variances by up to κ^2, but this would be a lot of extra effort, and we are hoping that κ^2 is small.

In summary, the inference would comprise many runs of the Kalman filter, to maximize $\log L(\theta)$ and find a plug-in value $\hat{\theta}$, and then one run of the Kalman smoother at $\theta = \hat{\theta}$, to compute the smoothing distributions of the latent process. The alternative to using a numerical optimizer on $\log L(\theta)$ is to use the EM algorithm; see Murphy (2012, sec. 17.5) or Cressie and Wikle (2011, sec. 8.3.1). This achieves potentially better performance of the optimizer (always traveling uphill) but at the cost of higher cost per iteration, because filtering and smoothing are required at every iteration.

4 | REFLECTIONS ON THE DSTM

Here are some initial reflections on the linear Gaussian DSTM in Section 3, to be clear about what it can and cannot do.

(1) It is a common feature of environmental datasets that observables can have very different spatial footprints, and the challenge of merging these, or making predictions for footprints different from those of the datasets, is termed the change of support problem (COSP, see, e.g., Cameletti, 2013). In the application in Section 2, the GPS observations are point-referenced, while the GRACE observations are areal averages. However, there is no COSP for a continuous representation of Y_t, as in (1), because a spatial integral is a linear transformation of Y_t. In Section 7 we will approximate spatial integrals with a numerical integration rule. This is straightforward in principle but, as discussed in Section 7, it can cause computational problems with massive datasets.

(2) The DSTM cannot handle observables which are emphatically discrete, such as Binomial, or Poisson with small mean. In this case the Gaussian approximation to the observables is hopeless. A more bespoke treatment is required, such as the package FRKv2 (Sainsbury-Dale et al., 2021) in the R statistical computing environment (R Core Team, 2020), which uses a Laplace approximation for the observed data likelihood, and Monte Carlo simulation for prediction and uncertainty quantification.

Sign-constrained observables are common in environmental science: typically observables that must be positive. These pose less of a problem for the DSTM. If the mean of these observables is well above zero, then a Gaussian distribution centered around the mean function $\bar{y}$ will work reasonably well, because it will assign only a small probability to a negative value, and this can be ignored for the purposes of reporting a predictive mean and standard deviation. If the probability of negative values is too large to ignore, then modeling in logarithms is a possibility. However, this breaks down for areal observables, because the integration needs to be over Y_t, not $\log(Y_t)$. In this case, a more general nonlinear model is required (e.g., Katzfuss et al., 2020). Cressie (2006) is a useful reference for modeling in logarithms.

Finally, there will be latent processes whose spatio-temporal dynamics are too complex for the Markov structure of the DSTM state equation, in which case the DSTM will fit poorly and its predictions will be unreliable. This might be a single latent process which the client wants to model at high resolution, or it might be two or more latent processes which couple in a complicated way, like temperature and precipitation. But the DSTM is not useless in this situation. First, if the mean function $\bar{y}$ is a dynamical simulation from a computer model of the latent processes, then the DSTM can be used to model limitations in the simulator, which will be systematic in space and time (Sha et al., 2019). Second, the state equation in the DSTM can be replaced by a simulator-based dynamical model, as in the Ensemble Kalman filter or the Unscented Kalman filter (Evensen, 2007; Julier & Uhlmann, 1997; Katzfuss et al., 2016). In this case, for massive datasets it will still be helpful for computational efficiency to consider the Sequential Updating condition, discussed next.

(3) There is a crucial condition in the DSTM which enables inference with datasets which are "massive", by which we mean comprising many thousands of observations every time-step. These massive datasets are common in environmental science, for example from ground-based stations on large spatial domains, like the GPS network in North America, or from remote sensing.

Let (Z_t^1, Z_t^2) be a partition of the observables Z_t. A standard and simple result states that if Z_t^1 and Z_t^2 are conditionally independent given α_t, then the Bayesian update of α_t by Z_t can proceed sequentially, first by Z_t^1 and then by Z_t^2. The one-line proof is (suppressing θ)

$$
\begin{aligned}
p(\alpha_t \mid z_t) &= \frac{p(z_t \mid \alpha_t) \cdot p(\alpha_t)}{p(z_t)} \\
&= \frac{p(z_t^1 \mid \alpha_t) \cdot p(z_t^2 \mid \alpha_t) \cdot p(\alpha_t)}{p(z_t^2 \mid z_t^1) \cdot p(z_t^1)} \\
&= \frac{p(z_t^2 \mid \alpha_t) \cdot p(\alpha_t \mid z_t^1)}{p(z_t^2 \mid z_t^1)}.
\end{aligned}
\tag{12}
$$

If applicable, this would reduce the computational cost of one time-step from $O\left((n_1 + n_2)^3\right)$ to $O\left(n_1^3 + n_2^3\right)$. On a current desktop, $n \sim 10^4$ is on the edge of computability, while $n_1, n_2 \sim 5 \times 10^3$ is very doable. Obviously, if Z_t can be partitioned into more conditionally independent components, the reduction in computational cost is even larger.

A similar condition is used in the "spatial partitioning" approach for Gaussian Process models for massive spatial datasets (see, e.g., Heaton et al., 2018), although in this case the conditional independence structure is used to factor the likelihood, so that the calculation of the log-likelihood can be dispatched in parallel over several nodes.

For the linear Gaussian DSTM with the parameters fixed, the conditional independence property is equivalent to the covariance property $\text{Cov}(Z_t^i, Z_t^j \mid \alpha_t) = 0$ where i and j are different components of a partition of Z_t. Since $\text{Var}(Z_t \mid \alpha_t) = \text{Var}(H_t \nu_t + \epsilon_t)$, the crucial condition for sequential updating in the DSTM is:

Definition 1 (Sequential updating, SU). Assuming that Σ_t is diagonal, sequential updating is available at time-step t whenever $\text{Var}(H_t \nu_t)$ is block-diagonal, with each block of observables being processed in turn. In the case where $\text{Var}(H_t \nu_t)$ is diagonal, any computationally-efficient arrangement of blocks is available.

Sequential updating in the Kalman filter is efficiently implemented using the canonical parameterization of the Multivariate Gaussian distribution; see Rue and Held (2005, ch. 2). The Kalman filter in the canonical parameterization is sometimes termed the "information filter" (Murphy, 2012, ch. 18).

(4) The SU condition indicates that the truncation error ν is a source of difficulty, as will be illustrated below. But dropping the truncation error (i.e., imposing $\kappa = 0$) should be avoided except *in extremis*. We do not expect to be able to capture Y in a basis expansion of our choosing, and may sometimes be forced to limit the number of basis functions for computational efficiency. So a truncation error of some sort is nearly inevitable, and we have to learn to work with it.

When Σ_t is diagonal, the SU condition requires that $H_t C_t H_t^T$ is block-diagonal, where C_t is the Gram matrix of the correlation function C at time-step t. This is more likely to occur if C_t is diagonal, or, failing that, sparse. So the correlation function C ought to have compact support to promote sparsity. The default choice for an isotropic C might be the spherical correlation function

$$C(s, s') = \psi\left(\frac{\|s - s'\|}{\ell}\right), \tag{13a}$$

where

$$\psi(x) := \begin{cases} 1 - (3/2)x + (1/2)x^3 & 0 \leq x \leq 1, \\ 0 & x > 1, \end{cases} \tag{13b}$$

see Gneiting (2002) for this and other options. Another approach would be to take a more general correlation function and "taper" it (Kaufman et al., 2008). This allows for more generality, including control over anisotropy. Below, ℓ will denote the correlation length of C, the value such that $C(s, s') \approx 0$ for all s, s' which are at least ℓ apart. While ℓ could be treated as an uncertain parameter in a specified C, it is more efficient to specify ℓ, and then to reorganize the observation equation, as discussed in Section 6 and Section 7.

One beneficial side-effect of including a truncation error with an unknown scale κ is that it can absorb observation error. On balance, the reported observation errors are more likely to understate than overstate the uncertainty induced by preprocessing and its assumptions. If Σ_t quantifies the reported observation error, then extra error in the observables has to go somewhere, and ν_t is the natural receptacle, having a similar (but not identical) spatio-temporal structure. Therefore in practice it may be better to think of κ as a lumped parameter, even though adjusting for under-reported observation error is not its primary purpose. But if $\kappa = 0$ has been imposed, it would be prudent to introduce a new parameter to scale Σ_t in the observation equation.

(5) Finally, there is no reason to think that M in the state equation will be sparse. Cressie and Wikle (2011, ch. 7) describe models in which it is natural for M to be dense, although it might have a low-dimensional parameterization. More generally, α_{tj} is the coefficient of the basis function ϕ_j and we fully expect different basis functions to have different temporal behavior. For example, ϕ might be a multiresolution basis, and perhaps the local basis functions behave differently to the global basis functions. Or else ϕ might be localized tent functions, and perhaps high latitude basis functions behave differently to low-latitude basis functions. Or else ϕ might capture physical features like coastlines, plains, mountain ranges, or river basins, and perhaps plains behave differently to mountains.

There is a huge computational advantage to the simplest representation $M = \rho I$ for some $0 < \rho < 1$, which follows from the Kronecker product structure of $\mathrm{Var}(\boldsymbol{\alpha})$, where

$$\boldsymbol{\alpha} := \begin{bmatrix} \boldsymbol{\alpha}_0 \\ \vdots \\ \boldsymbol{\alpha}_T \end{bmatrix}.$$

If $M = \rho I$, then $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, R \otimes Q)$, where R is the AR1 variance matrix with elements $R_{ij} = \rho^{|i-j|}/(1 - \rho^2)$. R^{-1} is sparse (tridiagonal), and this would lead to an efficient calculation if Q^{-1} was also sparse, as it would be if an SPDE approach is used (Lindgren et al., 2011), and every Σ_t is diagonal. Then the INLA approach can be applied to integrate out the parameters (Rue et al., 2009). This highly attractive package is described and used in Cameletti et al. (2013), to model daily particulate matter concentrations in the Piemonte region of northern Italy.

However, $M = \rho I$ is unlikely to be widely applicable in environmental science, as already discussed. For example, in the application in Section 2, the basis functions might be localized tent functions and the spatial domain might be divided into k large-scale river basins, for which

$$M = \mathrm{bdiag}(\rho_1 I_1, \dots, \rho_k I_k),$$

where "bdiag" denotes "block diagonal", $\rho_1, \dots, \rho_k$ are parameters to be estimated, one for each basin, and $I_1, \dots, I_k$ are identity matrices, one for each basin, after the basis functions have been ordered by basin. This breaks the Kronecker product structure for $\mathrm{Var}(\boldsymbol{\alpha})$. So M does not have to be complicated, but it is restrictive to assume that $M = \rho I$.

5 | HIGH RESOLUTION IN TIME

There will be applications where the observations have higher resolution than the DSTM in Section 3 can cope with. This section and the next explore two common cases. First, the observations may have higher temporal resolution than the time-step of the state equation, in this section. Second, the observations may be closer together in space than the correlation length of the truncation error, which violates the SU condition (Definition 1), in Section 6. In both cases one solution is to process the observables into "quasi-observables" that play the role of $\boldsymbol{Z}_t$ in the observation equation. Quasi-observables are distinguished from actual observables by writing $\tilde{\boldsymbol{Z}}_t$ rather than $\boldsymbol{Z}_t$.

This section and Section 6.2 are examples of "upscaling", in which high resolution observables ($\boldsymbol{Z}_t$) are used to create low-resolution summaries ($\tilde{\boldsymbol{Z}}_t$); this section is upscaling in time, and Section 6.2 is upscaling in space.

Consider the case where Y is continuous in time. To reduce clutter, fix a location s and write $Y(\tau) := Y(s, \tau)$, where τ is the continuous index of time. Despite the underlying process being continuous, the statistician has decided to model $Y(\cdot)$ discretely, say annually, usually for computational efficiency but also, perhaps, to filter out higher-frequency effects which are not of interest, including seasonality. This requires an aggregation function to map $Y[t, t + 1)$ to Y_t, adopting the "calendar" convention that year t comprises the time interval $[t, t + 1)$. There are two natural choices:

$$Y_t = Y(t) \qquad\qquad \text{annual start,} \qquad\qquad (14\text{a})$$

$$Y_t = \int_t^{t+1} Y(\tau)\, \mathrm{d}\tau \qquad\qquad \text{annual mean.} \qquad\qquad (14\text{b})$$

In the application in Section 2, GPS observations are available at a specified set of locations, roughly daily, although there are intervals of drop-out. Some temporal aggregation is unavoidable, because the computation cost of daily time-steps is too high. If the client only needs elevation annually, then for computational efficiency it seems natural to aggregate the GPS observations to annual time-steps.

Observations on the time-continuous process $Y(\cdot)$ at location s are made over the interval $[0, T + 1)$, of the general form

$$Z_i = Y(\tau_i) + \epsilon_i, \quad \epsilon_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_i^2), \quad i = 1, \dots, n. \qquad\qquad (15)$$

Here is one approach for mapping these observables to the annual summary Y_t, for the case where Y_t is the annual mean; the modification for the case where Y_t is the annual start is straightforward.

First, specify an interval $\mathcal{W}_t$ containing $[t, t+1)$, and suppose that within this window the process $Y(\cdot)$ has a temporal basis representation

$$Y(\tau) = \sum_{j=0}^{k} g_j(\tau) \cdot \beta_{tj} =: \mathbf{g}(\tau)^T \boldsymbol{\beta}_t, \quad \tau \in \mathcal{W}_t, \tag{16}$$

where $\mathbf{g}$ is a specified finite set of basis functions, and $\boldsymbol{\beta}_t$ are unknown fixed basis coefficients which belong to the window $\mathcal{W}_t$. Adding a truncation term to (16) would be superfluous, because we expect to fit $Y(\cdot)$ well within the window $\mathcal{W}_t$, and any truncation error will be dominated by observation error. Now use the observations to fit the estimator $\hat{\boldsymbol{\beta}}_t$, for which

$$\hat{\boldsymbol{\beta}}_t(\mathbf{Z}_{\mathcal{W}_t}) \overset{\text{app}}{\sim} \mathcal{N}(\boldsymbol{\beta}_t, \hat{V}_t), \quad \mathbf{Z}_{\mathcal{W}_t} := \{Z_i : \tau_i \in \mathcal{W}_t\}, \tag{17}$$

where "$\overset{\text{app}}{\sim}$" denotes "approximately distributed as" and $\hat{V}_t$ is the estimated sampling variance. If the observation errors σ_i in (15) are not given, or are unreliable, then $\hat{\boldsymbol{\beta}}_t$ can be fitted by ordinary least squares (OLS); otherwise, by generalized least squares (GLS). GLS can also be used if the ϵ_i are not independent, although it would be unusual for observations to be reported along with a nondiagonal variance matrix for the errors. Either way, $\hat{V}_t$ will scale with the observation errors $\{\sigma_i : \tau_i \in \mathcal{W}_t\}$.

Define

$$\bar{g}_j := \int_t^{t+1} g_j(\tau) \, d\tau, \quad j = 0, \dots, k, \tag{18a}$$

$$\tilde{Z}_t(\mathbf{Z}_{\mathcal{W}_t}) := \bar{\mathbf{g}}^T \hat{\boldsymbol{\beta}}_t(\mathbf{Z}_{\mathcal{W}_t}), \tag{18b}$$

where $\bar{\mathbf{g}}$ is the vector of $\bar{g}_j$ values. $\tilde{Z}_t$ will be the quasi-observable, whose direct observation equation is derived as follows. Suppress the '$(\mathbf{Z}_{\mathcal{W}_t})$' argument on $\tilde{Z}_t$ and $\hat{\boldsymbol{\beta}}_t$, to reduce clutter. Then

$$\begin{aligned} \tilde{Z}_t &= \bar{\mathbf{g}}^T \hat{\boldsymbol{\beta}}_t \\ &= \bar{\mathbf{g}}^T(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t + \boldsymbol{\beta}_t) \\ &= \bar{\mathbf{g}}^T(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t) + \bar{\mathbf{g}}^T \boldsymbol{\beta}_t \\ &= \bar{\mathbf{g}}^T(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t) + Y_t, \end{aligned} \tag{19}$$

where the last equality follows from (14b) and (16):

$$Y_t = \int_t^{t+1} \sum_j g_j(\tau) \cdot \beta_{tj} \, d\tau = \sum_j \int_t^{t+1} g_j(\tau) \, d\tau \cdot \beta_{tj} = \bar{\mathbf{g}}^T \boldsymbol{\beta}_t. \tag{20}$$

Then (17) and (19) imply that $\tilde{Z}_t \mid Y_t$ is approximately Normal, with

$$E(\tilde{Z}_t \mid Y_t) \approx Y_t, \quad \text{Var}(\tilde{Z}_t \mid Y_t) \approx \bar{\mathbf{g}}^T \hat{V}_t \, \bar{\mathbf{g}}. \tag{21}$$

The direct observation equation for $\tilde{Z}_t$ becomes

$$\tilde{Z}_t = Y_t + \tilde{\epsilon}_t, \quad \tilde{\epsilon}_t \sim \mathcal{N}(0, \bar{\mathbf{g}}^T \hat{V}_t \, \bar{\mathbf{g}}), \tag{22}$$

where the approximation error has been buried in the quasi-error term $\tilde{\epsilon}_t$.

Thus, at a given location, the time-series of observations is processed into a sequence of estimated $\{\hat{\boldsymbol{\beta}}_t, \hat{V}_t\}$ values, from which $\tilde{z}_t^{\text{obs}} = \bar{\mathbf{g}}^T \hat{\boldsymbol{\beta}}_t$, with the direct observation equation (22).
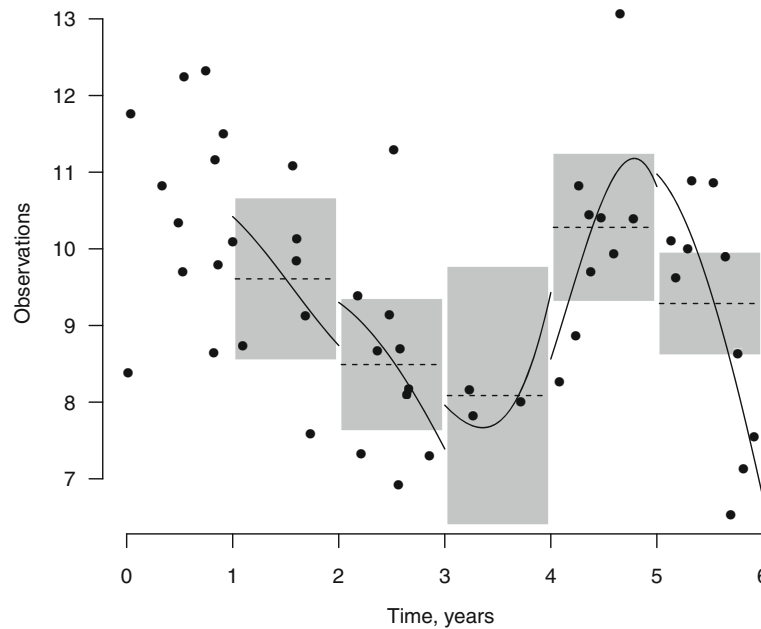
FIGURE 1 Illustration of the approach for aggregating observations that are high-resolution in time, using synthetic data. In this case, the window for $[t, t+1)$ is $\mathcal{W}_t = [t-1, t+1)$, and the basis functions are the first four Legendre polynomials (i.e., up to cubic). The dots show z_t^{obs}, the solid lines show the fitted polynomials using OLS, the dashed lines show the quasi-observation $\tilde{z}_t^{obs}$ using annual means, and the grey polygons show ± 2 standard deviations of the quasi-observation observation error.

The safest choice of window for $\tilde{Z}_t$ is $\mathcal{W}_t = [t, t+1)$, because in this case the observation error on each Z_i occurs in exactly one $\tilde{Z}_t$. Otherwise the observation errors will be replicated in consecutive $\tilde{Z}_t$, which violates the property that ϵ_t is independent across time. But this small covariance is likely to be ignorable. In the absence of better judgments, one option is to use the window $\mathcal{W}_t = [t-1, t+1)$ for $\tilde{Z}_t$, and the basis functions

$$g_j(\tau) = P_j(\tau - t), \quad \tau \in [t-1, t+1), \tag{23}$$

where P_j is the jth order Legendre polynomial (see, e.g., Kreyszig, 1978, sec. 3.7.1). These basis functions have the mild advantage of producing a roughly orthogonal model matrix for OLS when the observations are roughly evenly-distributed through $[t-1, t+1)$. A cubic, for which $j = 0, \ldots, 3$, will often be sufficient to capture the behavior of $Y(\cdot)$ in this narrow window. Figure 1 illustrates this approach using synthetic data.

Both the window and the set of basis functions can vary with t at the same location, and if there are few or no observations in an interval $[t, t+1)$ then $\tilde{Z}_t$ can just be dropped. If there is seasonality in the time-series, it can be absorbed by additional seasonal basis functions added to (16). The main effect of these is to reduce the size of $\hat{V}_t$; that is, to prevent seasonality from contributing, wrongly, to uncertainty about $\tilde{Z}_t$.

The generalization of these quasi-observations to multiple locations is straightforward, with each location being processed separately, and then stacked together into each $\tilde{Z}_t$. Overall, high resolution in time at fixed locations does not seem to cause any difficulties in the DSTM, and nearly all of the observations get used, albeit after processing into a reduced set of quasi-observations.

Unfortunately, this upscaling approach does not generalize to the case where the observations are scattered in both space and time, like observations from ocean drifters, which report with a high frequency in time, but from a different location each time. One superficially attractive solution would be to shorten the time-step until nearly every observation can be mapped to the start of a time-step. This increases the computational cost of filtering and smoothing, but only linearly. However, the first-order dynamical model for $\boldsymbol{\alpha}_t$, which might seem natural on a time-step of a year, might be much less appropriate for a time-step of, say, a week. A higher-order dynamical model for $\boldsymbol{\alpha}_t$ would require many more components in the state vector, and this would substantially increase the cost of smoothing. Modeling ocean drifter

observations at their natural resolution is very demanding, and one current approach uses nonlinear methods (particle filters) and high-performance computing (Holm et al., 2020).

6 | HIGH RESOLUTION IN SPACE

In the application in Section 2, GPS locations are not uniformly distributed across the domain: in some regions they are highly concentrated, with separations of just a few kilometers. This means that $C(s, s')$ is not approximately zero for many pairs of locations, where C is the correlation function of ν, the truncation error. Therefore there will be a clash between including the truncation error, and meeting the SU condition; see Section 4.

In basic terms, there are three possible cases:

1. The dataset is not massive, in which case the SU condition is not required for computational feasibility, although it will still provide computational efficiency.
2. The dataset is massive, but the truncation error is ignored (i.e., $\kappa = 0$). In this case the SU condition holds automatically, under the assumption that Σ_t is diagonal.
3. The dataset is massive, and the truncation error is included. In this case the dataset must be preprocessed in some way for computational feasibility.

Here, we focus on the third case. There are two simple solutions: "subsampling" and "aggregating". Both of them require a specified spatial length, denoted ℓ, the correlation length of C. Aggregating also requires a specified form for C, such as the spherical correlation function in (13).

In both solutions, the net effect is to discard some of the information in the dataset, for computational feasibility and efficiency. From another point of view, we are addressing the question: if it is necessary to discard information in order to attain computational feasibility/efficiency, what is a principled way to proceed, which takes account of the structure of the DSTM? Our answer is that it depends on the correlation length of the truncation error.

6.1 | Subsampling

The simplest solution is to selectively subsample the locations, until all distances exceed the spatial correlation length ℓ. Subsampling can proceed deterministically, once a correlation length has been specified. First, discard locations with dubious observations, or for which there are likely to be unmodeled influences. Then proceed sequentially, at each stage taking the pair of locations with the smallest separation, and deleting the one with the larger observation error. Terminate when the smallest separation exceeds the correlation length.

Deterministic subsampling can be applied separately to each time-step, but if the locations are the same or nearly the same for every time-step, then subsampling may permanently exclude some locations. An alternative is to subsample stochastically, which produces a different subset of locations for each time-step, and has the effect of ensuring that most locations get used at least once (which could be enforced, if required). The "kmeans++" algorithm is one possibility (Arthur & Vassilvitskii, 2007). One location is chosen uniformly at random as the first location. The second location is chosen at random using probabilities proportional to the square of the distance to the first location. The algorithm proceeds sequentially, using distance to the nearest already-selected location. When subsampling locations, this algorithm should be modified to set the probability of selection for locations within ℓ of the already-selected locations to zero. Figure 2 illustrates this approach.

Let $\tilde{Z}_t$ be the subsampled set of observables, at locations $s_1, \ldots, s_m$. The indirect observation equation is

$$\tilde{Z}_t = (\bar{y}_t + \Phi_t \alpha_t) + \gamma_t, \quad \gamma_t \overset{\text{ind}}{\sim} \mathcal{N}(0, R_t), \tag{24a}$$

where

$$R_t = \kappa^2 I + \tilde{\Sigma}_t, \tag{24b}$$

where $\tilde{\Sigma}_t$ is the observation error variance Σ_t subsampled from Z_t to $\tilde{Z}_t$. The first term in R_t is $\text{Var}(\nu_t)$ when all of the locations are separated by at least ℓ. If Σ_t is diagonal, then R_t is diagonal. The SU condition is satisfied, and the
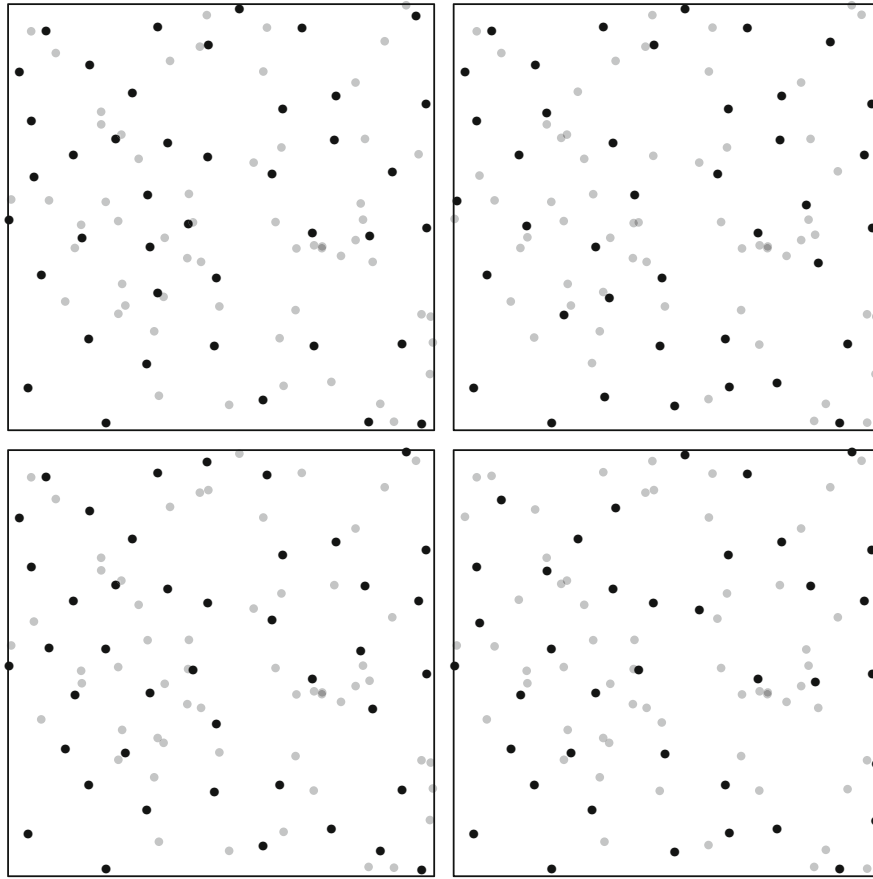
FIGURE 2 Four stochastic subsamplings of the same 100 points distributed uniformly at random in the unit cube, with a correlation length of $\ell = 0.1$, using a modification of the "kmeans++" algorithm.

Kalman filter update at time-step t can be performed sequentially, in whatever arrangement of observations is most efficient.

6.2 | Aggregating

In "aggregating", the spatial domain is tiled using n polygonal tiles each of which are at least ℓ in their minimum width. Within each tile, the observables are merged into an arithmetic mean, which becomes the quasi-observable for the tile. These quasi-observables are treated as if their locations were at least ℓ apart. This is spatial upscaling, with the objective of satisfying the SU condition.

Here are the details; the critical assumption is at (27). Let

$$\boldsymbol{Z}_t^i := \left(Z_{t1}^i, \ldots, Z_{tk_i}^i \right)$$

be the k_i observables within tile i, with locations $s_1^i, \ldots, s_{k_i}^i$. Let $s_1, \ldots, s_m$ be the full set of all locations for all n tiles, ordered by tile, and let $\boldsymbol{Z}_t$ and $\boldsymbol{Y}_t$ be the corresponding vector of observables and latent process values. The quasi-observable for tile i is the arithmetic mean of the observables in tile i, which gives the direct observation equation

$$\tilde{Z}_t^i = H_t^i \boldsymbol{Z}_t = H_t^i \left(\boldsymbol{Y}_t + \boldsymbol{\epsilon}_t \right), \quad i = 1, \ldots, n, \tag{25}$$

where the $1 \times m$ matrix H_t^i is $1/k_i$ for the locations in tile i, and zero elsewhere. Stack these n quasi-observables together to get the indirect observation equation

$$\tilde{\boldsymbol{Z}}_t = H_t(\bar{\boldsymbol{y}}_t + \Phi_t \boldsymbol{\alpha}_t + \boldsymbol{v}_t + \boldsymbol{\epsilon}_t), \tag{26a}$$

where

$$H_t := \begin{bmatrix} H_t^1 \\ \vdots \\ H_t^n \end{bmatrix}, \tag{26b}$$

and H_t is block-diagonal, because each location occurs in exactly one tile.

The simplifying assumption is

$$\mathrm{Cov}\left(H_t^i v_t, H_t^j v_t\right) = \mathbf{0}, \quad \text{for } i \neq j, \tag{27}$$

to a good approximation. In other words, the arithmetic means of the truncation errors within tiles of minimum width ℓ are uncorrelated across tiles (see below, and Figure 3). This implies that

$$\mathrm{Var}(H_t v_t) = \kappa^2 \cdot \mathrm{diag}\left((k_1)^{-2}\mathbf{1}^T C_t^1 \mathbf{1}, \ \dots \ , (k_n)^{-2}\mathbf{1}^T C_t^n \mathbf{1}\right) =: C_t, \tag{28}$$

where C_t^i is the $k_i \times k_i$ Gram matrix for C at the locations $s_1^i, \dots, s_{k_i}^i$. The indirect observation equation becomes

$$\tilde{Z}_t = H_t\left(\overline{y}_t + \Phi_t \alpha_t\right) + \gamma_t, \quad \gamma_t \overset{\mathrm{ind}}{\sim} \mathcal{N}(\mathbf{0}, R_t) \tag{29a}$$

where

$$R_t := C_t + H_t \Sigma_t H_t^T, \tag{29b}$$

and C_t is diagonal. If Σ_t is diagonal, then the block-diagonal structure of H_t implies that $H_t \Sigma_t H_t^T$ is diagonal, so that R_t is diagonal. Just as in subsampling, the SU condition is satisfied, and the Kalman filter update at time-step t can be performed sequentially.
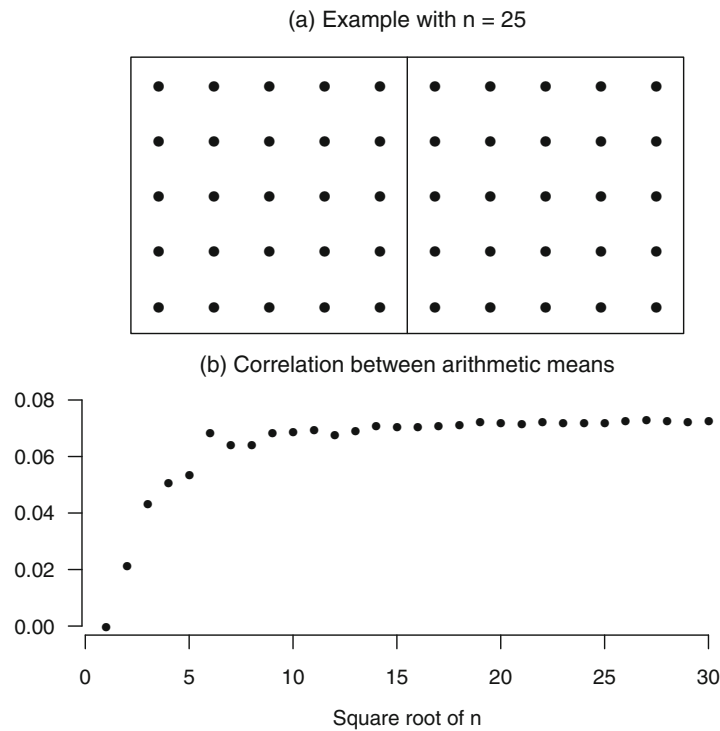

(a) Example with n = 25

(b) Correlation between arithmetic means

Square root of n

FIGURE 3 Two contiguous square tiles with edge length ℓ. The top panel shows the layout of $n = 25$ evenly-spaced locations for each tile. The bottom panel shows the correlation between the arithmetic means of the values at the locations from the two tiles, using the spherical correlation function, (13). The calculation is exact, and the irregularity in the bottom panel reflects the discrete nature of the grid of locations, interacting with the compact support of the correlation function.

A back-of-the-envelope calculation suggests that tiles with minimum width ℓ suffice, based on the spherical correlation function in (13). Figure 3 shows the correlation between two contiguous square tiles with evenly-spaced locations, for different numbers of locations. Although there will be some configurations of tiles and locations where the correlation is high (e.g., where the edge of the tile goes through an isolated clump of locations), these ought to be avoidable because the tiling is adjustable. According to Figure 3, applying (27) with tile widths of at least ℓ will zero-out correlations in the quasi-observables which are not more than about 0.07.

In summary, aggregating requires a specified correlation length, a specified tiling of the spatial domain, and a specified correlation function, introducing quantities about which the domain experts may have only weak judgments. Subsampling, on the other hand, requires just a correlation length, used qualitatively. From this point of view, subsampling seems less intrusive, from a modeling point of view, if the client can tolerate using only a subset of the observations. On the other hand, if the observation errors are not quite Gaussian, then aggregation has the advantage of tending to make the pseudo-observable observation errors more Gaussian, through the same mechanism as the Central Limit Theorem (see, e.g., Grimmett & Stirzaker, 2001, sec. 5.10).

7 | AREAL OBSERVATIONS

In the case of observations which are high resolution in space, subsampling and aggregating are two approaches to satisfying the SU condition; see Section 6. Subsampling side-stepped the form of C, the correlation function of truncation error, while aggregating did not. So at least there was the option of side-stepping C. Unfortunately, with areal observations there is no side-step, except for the extreme one of assuming that the truncation error can be ignored (i.e., $\kappa = 0$).

Consider an areal observable with the definition

$$
\begin{aligned}
Z_{ti} &:= \frac{1}{|\mathcal{F}_i|} \int_{\mathcal{F}_i} Y_t(s) \, \mathrm{d}s + \epsilon_{ti} \\
&= \frac{1}{|\mathcal{F}_i|} \int_{\mathcal{F}_i} \left\{ \bar{y}_t(s) + \boldsymbol{\phi}(s)^T \boldsymbol{\alpha}_t + v_t(s) \right\} \, \mathrm{d}s + \epsilon_{ti},
\end{aligned}
\tag{30}
$$

where $\mathcal{F}_i$ is the "footprint" of the ith observable, assumed to be the same for every time-step, to reduce clutter.

In principle, the space-integral can be taken over each of $\bar{y}_t$, $\boldsymbol{\phi}$ and v_t, to create an exact indirect observation equation for Z_{ti}, as proposed by Fuentes and Raftery (2005). But this is computationally demanding, and likely to involve approximations if any of these three functions can only be evaluated pointwise. So instead the space-integral is replaced by a numerical integration rule, for which a midpoint rule is the simplest implementation.

Therefore, let $\mathcal{G}$ be an axially-aligned grid with horizontal and vertical spacing d, which covers $\mathcal{F}_i$, and let $\mathcal{H}_i = \mathcal{F}_i \cap \mathcal{G}$. Then, provided that $\mathcal{G}$ is sufficiently fine (i.e., d is sufficiently small),

$$
\int_{\mathcal{F}_i} Y_t(s) \, \mathrm{d}s = d^2 \sum_{s \in \mathcal{H}_i} Y_t(s), \quad |\mathcal{F}_i| = d^2 |\mathcal{H}_i| = d^2 k_i,
\tag{31}
$$

to a good approximation, where k_i is the number of gridpoints in $\mathcal{H}_i$. Hence,

$$
Z_{ti} = H_t^i \boldsymbol{Y}_t^i + \epsilon_{ti}
\tag{32a}
$$

where

$$
H_t^i := k_i^{-1} \mathbf{1}^T, \quad \boldsymbol{Y}_t^i := \begin{bmatrix} Y_t(s_1^i) \\ \vdots \\ Y_t(s_{k_i}^i) \end{bmatrix},
\tag{32b}
$$

and $s_1^i, \ldots, s_{k_i}^i$ are the grid locations in $\mathcal{H}_i$. Stack these together to get the direct observation equation

$$\boldsymbol{Z}_t = H_t \boldsymbol{Y}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{0}, \Sigma_t), \tag{33a}$$

where

$$\boldsymbol{Z}_t := \begin{bmatrix} Z_{t1} \\ \vdots \\ Z_{tn} \end{bmatrix}, \quad \boldsymbol{Y}_t := \begin{bmatrix} \boldsymbol{Y}_t^1 \\ \vdots \\ \boldsymbol{Y}_t^n \end{bmatrix} \tag{33b}$$

and

$$H_t := \text{bdiag}\left((k_1)^{-1}\mathbf{1}^T, \ldots, (k_n)^{-1}\mathbf{1}^T \right), \tag{33c}$$

where "bdiag" denotes "block diagonal". There is no requirement that the footprints are disjoint, and in fact it is a powerful feature of the DSTM that it can merge point and areal observations in complete generality, including overlaps.

Unfortunately, though, there may be trouble ahead. Let $s_1, \ldots, s_m$ be the full set of gridpoints in the direct observation equation at time-step t. The indirect observation equation is

$$\boldsymbol{Z}_t = H_t(\bar{\boldsymbol{y}}_t + \Phi_t \boldsymbol{\alpha}_t) + \boldsymbol{\gamma}_t, \quad \boldsymbol{\gamma}_t \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{0}, R_t), \tag{34a}$$

where

$$R_t := \kappa^2 H_t C_t H_t^T + \Sigma_t, \tag{34b}$$

and C_t is the Gram matrix of the correlation function C for $s_1, \ldots, s_m$. Unfortunately R_t is unlikely to be block-diagonal, especially if the footprints are small and contiguous, like the GRACE polygons from the application in Section 2. Therefore the SU condition is likely to be violated, in which case there is no sequential updating in the Kalman filter at time-step t.

There are several options to recover the SU condition with areal observations. First, if the footprints are disjoint and their minimum widths are typically larger than ℓ, the correlation length of C, then the off-diagonal blocks of $H_t C_t H_t^T$ could be zeroed-out, exactly as in Section 6.2. Second, if their minimum widths are typically smaller than ℓ, contiguous footprints could be merged in order to increase the minimum width to at least ℓ, and then apply the first option. This would be the natural solution for the GRACE polygons in the application in Section 2. Third, if merging is unpalatable to the domain experts, then the footprints could be subsampled, exactly as in Section 6.1. Of course subsampling might be unpalatable to the client. Finally, the truncation error might simply be ignored (i.e., $\kappa = 0$), which may not compromise the model if κ is small, and if an additional parameter is introduced to rescale Σ_t, as discussed in Section 4, item (4).

8 | LOCAL TRENDS

The only source of trends in the DSTM is the mean function $\bar{y}$: in the absence of shocks, $\boldsymbol{\alpha}_t$ tends back to $\boldsymbol{0}$ and Y_t tends back to $\bar{y}_t$. Most environmental processes contain trends, some of which may be extensive in space and enduring in time, such as the trends from climate change. But many will be local in space and limited in time. In the application in Section 2, elevation contains local trends, because it responds to processes which have durations and response times of several years, such as weather patterns like El Niño, which affect water storage in lakes, reservoirs, and aquifers. Ideally, $\bar{y}$ would contain covariates which represent these trends, but in practice that is challenging. Therefore we will often want to give the DSTM the flexibility to include local trends, which means, in effect, that if $Y_{t-2}(s) \le Y_{t-1}(s)$, then it is more probable that $Y_{t-1}(s') \le Y_t(s')$, for s' in some neighborhood of s.

8.1 | Local linear trend model

A natural way to introduce a local trend is to use the basis expansion on the *differences* in $Y_t(s)$:

$$Y_t(s) - Y_{t-1}(s) = \phi(s)^T \alpha_t + \nu_t(s), \tag{35}$$

where the mean function $\bar{y}$ has now canceled out (but will reappear below). (35) could also include a mean function for the differences, which does not create any complications, and so we have dropped it to reduce clutter. In the application in Section 2, GRACE includes a contribution from glacial isostatic adjustment (GIA), an annual change which is effectively constant in time over a decade or so, but varies in space (Vishwakarma et al., 2022). So we might include a specified GIA mean function in (35), estimated from a simulation such as ICE6G (Peltier et al., 2015). More generally, this mean function might include space- and time-varying covariates for the change in the latent process, with uncertain coefficients.

What is the implication of modeling the differences? Represent Y_0 as

$$Y_0(s) - \bar{y}_0(s) = \phi(s)^T \alpha_0 + \nu_0(s), \tag{36}$$

where the original mean function at time $t = 0$ has reappeared. Expanding out using (35),

$$\begin{aligned} Y_t(s) - \bar{y}_0(s) &= (Y_0(s) - \bar{y}_0(s)) + (Y_1(s) - Y_0(s)) + \cdots + (Y_t(s) - Y_{t-1}(s)) \\ &= \phi(s)^T \alpha_0 + \cdots + \phi(s)^T \alpha_t + \nu_0(s) + \cdots + \nu_t(s) \\ &= \phi(s)^T (\beta_t + \alpha_t) + \nu_0(s) + \cdots + \nu_t(s), \end{aligned} \tag{37}$$

where $\beta_0 := 0$, and $\beta_t := \beta_{t-1} + \alpha_{t-1}$. Therefore, the original state equation (6) must be augmented to allow β_t to accumulate α_{t-1}:

$$\begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix} = \begin{bmatrix} M & 0 \\ I & I \end{bmatrix} \begin{bmatrix} \alpha_{t-1} \\ \beta_{t-1} \end{bmatrix} + \begin{bmatrix} \eta_t \\ 0 \end{bmatrix}, \quad \eta_t \overset{\text{ind}}{\sim} \mathcal{N}(0, Q), \tag{38}$$

where the role of M and Q are not the same as in Section 3, because they apply to differences of Y, not to Y itself. The indirect observation equation, (11), has to be changed to

$$Z_t = H_t [\bar{y}_0 + \Phi_t(\alpha_t + \beta_t)] + \gamma_t, \quad \gamma_t \overset{\text{ind}}{\sim} \mathcal{N}(0, R_t), \tag{39a}$$

where $\gamma_t := H_t(\nu_0 + \cdots + \nu_t) + \epsilon_t$, and so

$$R_t := (t+1)\kappa^2 H_t C_t H_t^T + \Sigma_t, \tag{39b}$$

where C_t is the Gram matrix of C for the locations at time t, and making the convenient (but suspect) assumption that $\text{Var}(\nu_0(s)) = \kappa^2$. The state equation (38) might also include a stochastic term for β_t, represented as a nonzero variance matrix which would be added to the parameters.

This approach is similar to using a *local linear trend* in state-space modeling (see, e.g., Harvey, 1989). It is the "right" way to include a local trend in Y_t, because it does not change the underlying structure of the model, and so any method which has been developed to do inference about the DSTM can still be used, including the upscaling methods of Section 5 and Section 6. However, it doubles the length of the state vector, from α_t to (α_t, β_t). As discussed in Section 3.2, this increases the cost of smoothing eight-fold. So it is possible that this approach may be too expensive for some applications, in which case the computationally cheaper option in Section 8.2 might be useful.

8.2 | Differencing

Perhaps there is a quick-and-dirty way to implement local trends, without doubling the length of the state vector? Generally, there is not, but there is one special case, where the same set of observables is used at every time-step (to be generalized

below). In this case the direct observation equation is

$$\boldsymbol{Z}_t = H\boldsymbol{Y}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \stackrel{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{0}, \Sigma_t), \tag{40}$$

where $\boldsymbol{Y}_t$ and $\boldsymbol{Z}_t$ are always evaluated at the same specified locations. Differencing both sides of (40),

$$\begin{aligned}
\tilde{\boldsymbol{Z}}_t &:= \boldsymbol{Z}_t - \boldsymbol{Z}_{t-1} \\
&= H(\boldsymbol{Y}_t - \boldsymbol{Y}_{t-1}) + \boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_{t-1} \\
&= H(\Phi\boldsymbol{\alpha}_t + \boldsymbol{v}_t) + \boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_{t-1},
\end{aligned} \tag{41}$$

from (35). The $\boldsymbol{\beta}_t$ have dropped out of the indirect observation equation, and they are now superfluous in the state equation, and can be dropped there as well. Now the calculation is the same size as the DSTM in Section 3. Unfortunately, information about the level of the latent process has been lost. So differencing would only be a way to include local trends in the latent process if the client was uninterested in the level of the latent process, but only in the time-differences.

For completeness, note that differencing does not require that the incidence matrix H is time-invariant: differencing can be extended to keep more observations, but it is still necessary to discard every observable which does not have a corresponding value one time-step before.

There is one statistical *caveat* for differencing, concerning the error term in (41),

$$\boldsymbol{\gamma}_t := H\boldsymbol{v}_t + \boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_{t-1}. \tag{42}$$

The modeling assumption is that $\boldsymbol{\gamma}_t$ and $\boldsymbol{\gamma}_{t-1}$ are independent. But now they are dependent, because of the appearance of the same observation error in consecutive $\boldsymbol{\gamma}_t$:

$$\text{Cov}(\boldsymbol{\gamma}_t, \boldsymbol{\gamma}_{t-1}) = -\Sigma_{t-1} \neq \boldsymbol{0}. \tag{43}$$

So to summarize, *if* most observations are available at most time-steps, *and* the client does not care about the level of the latent process but only about the time-differences, *and* the observation error is small, then differencing the observables is a computationally efficient way to include local trends in the latent process, within the DSTM.

This is quite a restrictive set of conditions, especially in environmental science. The datasets for many environmental applications have limited duration, and have to be overlapped, or sometimes there is a gap of a year or more. Remote sensing products for example, change with the satellites which produce them (see, e.g., Schröder et al., 2019). Other datasets are attached to sensors which change location through time, like ocean drifters, mentioned in Section 5. Other datasets are collected with targeted campaigns at a specific time and location, like plane-based LIDAR. In these cases, differencing discards a lot of observations. Furthermore, many environmental processes are complicated and hard to measure, so we can expect the observation errors to be large (or, as discussed above, under-reported). Indeed, perhaps the most compelling reason for a statistical approach is precisely *because* the observations are sparse and time-varying, and the observation errors are large. Therefore the local linear trend approach in Section 8.1 should be favored over differencing, where possible.

8.3 | Observables which are differences

In the application in Section 2, the GPS observations are on the latent process, elevation, but the GRACE observations are on the time-differences of the latent process. The indirect observation equation in (11) cannot handle the GRACE observations, and therefore the DSTM in Section 3 is not appropriate for merging GPS and GRACE into an spatial-temporal process for elevation. Although, as discussed in Section 8.2, it may be suitable for merging GRACE and a subset of GPS into an updated spatial-temporal process for time-differences in elevation, after time-differencing the GPS observations.

The good news is that this issue has already been solved in the local linear trend model of Section 8.1, because time-differences in the latent process are explicitly modeled by $\boldsymbol{\alpha}_t$. So GPS observables, which measure $\boldsymbol{Y}_t$, are attached to $\boldsymbol{\alpha}_t + \boldsymbol{\beta}_t$ in the indirect observation equation, while GRACE observables, which measure $\boldsymbol{Y}_t - \boldsymbol{Y}_{t-1}$, are attached to $\boldsymbol{\alpha}_t$ in the indirect observation equation.

9 | CONCLUSION

The three challenges outlined in Section 1 were that environmental processes are complex, that environmental datasets are often complicated and irregular, and that environmental datasets are often massive. The interaction of these three challenges severely constrains the types of statistical model and inference than can be used in environmental science. We have shown that the "vanilla" Kalman filter, the DSTM outlined in Section 3, can, with careful implementation, handle these three challenges. That is, many spatio-temporal applications in environmental science are in the scope of the Kalman filter.

The argument in the paper runs as follows. The complexity of environmental processes implies that it is hard to capture their spatio-temporal behavior in a specified set of basis functions. Therefore we must expect a truncation error in the model which should not be treated in a simplistic fashion; for example, it should not be merged with an IID observation error in the observation equation. But if we recognize the possibility of structure in the truncation error, even in quite a primitive fashion, then we risk violating the Sequential Updating (SU) condition which is crucial for handling massive datasets, discussed in Section 4.

The central part of the article, Section 5 to Section 7, is about how to incorporate massive datasets without violating the SU condition. The biggest difficulty is with areal observations that are supplied on small contiguous polygons, like the GRACE dataset in our application in Section 2. In this case, and in other cases involving datasets which are high resolution in space, some form of subsampling or aggregation can recover the SU condition. These both involve discarding some of the information in the dataset, for computational feasibility and efficiency.

Finally, Section 8 is a generalization of the model in Section 3, which once again reflects the complexity of environmental processes. In many cases, the latent process (representing the quantities of interest) is driven by other processes, which can be included in the model as covariates. But these other processes may not have been measured, or even known. In this case, the dynamical behavior of the latent process will often be more complex than the model in Section 3 can capture. In the first instance, the latent process may have trends that are local in space or time. These trends can be accommodated by taking the model up a level, in a mathematical sense, going from a stochastic process on the levels to a stochastic process on the differences, known as a "local linear trend". However, this adjustment is expensive, because it doubles the number of coefficients, which increases the cost of the inference eightfold. So it is tempting to adjust the datasets rather than the model, by differencing. This may be the only option but, as we explain, it comes with *caveats*, and caution is advised.

DATA AVAILABILITY STATEMENT
Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID
Jonathan Rougier https://orcid.org/0000-0003-3072-7043
Aoibheann Brady https://orcid.org/0000-0003-0314-5144

REFERENCES
Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. *Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms*, 1027–1035.

Blewitt, G., Hammond, W., & Kreemer, C. (2018). Harnessing the GPS data explosion for interdisciplinary science. *Eos*, *99*. https://doi.org/10.1029/2018EO104623

Bolin, D., Wallin, J., & Lindgren, F. (2019). Latent Gaussian random field mixture models. *Computational Statistics and Data Analysis*, *130*, 80–93.

Cameletti, M. (2013). The change of support problem through the INLA approach. *Statistica & Applicazioni*, 29–43.

Cameletti, M., Lindgren, F., Simpson, D., & Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, *97*, 109–131.

Chao, B. F. (2016). Caveats on the equivalent water thickness and surface mascon solutions derived from the GRACE satellite-observed time-variable gravity. *Journal of Geodesy*, *90*, 807–813.

Cressie, N. (2006). Block kriging for lognormal spatial processes. *Mathematical Geology*, *38*(4), 413–443.

Cressie, N., & Wikle, C. K. (2011). *Statistics for Spatio-temporal data*. John Wiley & Sons, Inc.

Evensen, G. (2007). *Data assimilation: The ensemble Kalman filter*. Springer.

Fuentes, M., & Raftery, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics*, *61*, 36–45.

Gneiting, T. (2002). Compactly supported correlation functions. *Journal of Multivariate Analysis*, *83*, 493–508.

Grimmett, G. R., & Stirzaker, D. R. (2001). *Probability and random processes* (3rd ed.). Oxford University Press.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

He, X., Montillet, J.-P., Bos, M. S., Fernandes, R. M. S., Jiang, W., & Yu, K. (2020). *Filtering of GPS time series using geophysical models and common mode error analysis*. In J.-P. Montillet & M. S. Bos (Eds.), *Geodetic time series analysis in earth sciences* (pp. 261–273). Springer.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D. W., Sun, F., & Zammit-Mangion, A. (2018). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological, and Environmental Statistics*, *24*(3), 398–425.

Holm, H. H., Saetra, M. L., & Brodtkorb, A. R. (2020). *Data assimilation for ocean drift trajectories using massive ensembles and GPUs*. In R. Klöfkorn, E. Keilegavlen, F. A. Radu, & J. Fuhrmann (Eds.), *Finite volumes for complex applications IX: Methods, theoretical aspects, examples* Springer Proceedings in Mathematics & Statistics (Vol. *323*, pp. 715–724). Springer.

Julier, S. J., & Uhlmann, J. K. (1997). *New extension of the Kalman filter to nonlinear systems*. In I. Kadar (Ed.), *Signal processing, sensor fusion, and target recognition VI* (Vol. *3068*, pp. 182–193). SPIE: International Society for Optics and Photonics.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*, 35–45.

Kaplan, E. D., & Hegarty, C. (Eds.). (2017). *Understanding GPS/GNSS: Principles and applications*. Artech House.

Katzfuss, M., Stroud, J. R., & Wikle, C. K. (2016). Understanding the ensemble Kalman filter. *The American Statistician*, *70*(4), 350–357.

Katzfuss, M., Stroud, J. R., & Wikle, C. K. (2020). Ensemble Kalman filter methods for high-dimensional hierarchical dynamic space-time models. *Journal of the American Statistical Association*, *115*(530), 866–885.

Kaufman, C., Schervish, M., & Nychka, D. (2008). Covariance tapering for likelihood-based estimation in large spatial datasets. *Journal of the American Statistical Association*, *103*, 1556–1569.

Kreyszig, E. (1978). *Introductory functional analysis*. John Wiley & Sons.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series B*, *73*(4), 423–498.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. Harcourt Brace & Co.

Murphy, K. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Peltier, W. R., Argus, D. F., & Drummond, R. (2015). Space geodesy constrains ICE-age terminal deglaciation: The global ICE6G_C(VM5a) model. *Journal of Geophysical Research: Solid Earth*, *120*, 450–487.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing https://www.R-project.org/

Rue, H., & Held, L. (2005). *Gaussian Markov random fields: Theory and applications* Monographs on Statistics and Applied Probability (Vol. *104*). Chapman & Hall/CRC.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, *71*(2), 319–392 (with discussion).

Sainsbury-Dale, M., Zammit-Mangion, A., & Cressie, N. (2021). Modelling, fitting, and prediction with non-Gaussian spatial and spatio-temporal data using FRK. https://doi.org/10.48550/arXiv.2110.02507

Schröder, L., Horwath, M., Dietrich, R., Helm, V., van den Broeke, M. R., & Ligtenberg, S. R. M. (2019). Four decades of Antarctic surface elevation changes from multi-mission satellite altimetry. *The Cryosphere*, *13*, 427–449.

Sha, Z., Rougier, J. C., Schumacher, M., & Bamber, J. L. (2019). Bayesian model-data synthesis with an application to global Glacio-isostatic adjustment. *Environmetrics*, *30*(1), e2530.

Vishwakarma, B. D., Howarth, M., Groh, A., & Bamber, J. L. (2022). Accounting for GIA signal in GRACE products. *Geophysical Journal International*, *228*(3), 2056–2060.

Watkins, M. M., Wiese, D. N., Yuan, D.-N., Boening, C., & Landerer, F. W. (2015). Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons. *Journal of Geophysical Research: Solid Earth*, *120*(4), 2648–2671.

Wendland, H. (2018). *Numerical linear algebra: An introduction*. Cambridge University Press.