A Dimension-Reduced Approach to Space-Time Kalman Filtering

Christopher K. Wikle; Noel Cressie


Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28199912%2986%3A4%3C815%3AADATS%3E2.0.CO%3B2-Y

Biometrika is currently published by Biometrika Trust.

Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/bio.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.
A dimension-reduced approach to space-time Kalman filtering

BY CHRISTOPHER K. WIKLE
Department of Statistics, University of Missouri-Columbia, Columbia, Missouri 65211, U.S.A.
wikle@stat.missouri.edu

AND NOEL CRESSIE
Department of Statistics, The Ohio State University, Columbus, Ohio 43210, U.S.A.
ncressie@stat.ohio-state.edu

SUMMARY

Many physical/biological processes involve variability over both space and time. As a result of difficulties caused by large datasets and the modelling of space, time and spatio-temporal interactions, traditional space-time methods are limited. In this paper, we present an approach to space-time prediction that achieves dimension reduction and uses a statistical model that is temporally dynamic and spatially descriptive. That is, it exploits the unidirectional flow of time, in an autoregressive framework, and is spatially ‘descriptive’ in that the autoregressive process is spatially coloured. With the inclusion of a measurement equation, this formulation naturally leads to the development of a spatio-temporal Kalman filter that achieves dimension reduction in the analysis of large spatio-temporal datasets. Unlike other recent space-time Kalman filters, our model also allows a non-dynamic spatial component. The method is applied to a dataset of near-surface winds over the tropical Pacific ocean. Spatial predictions with this dataset are improved by considering the additional non-dynamic spatial process. The improvement becomes more pronounced as the signal-to-noise ratio decreases.

Some key words: Dynamic model; Empirical Bayes; Empirical orthogonal functions; Kriging; Large dataset; Optimal interpolation; Spatio-temporal modelling; Wind.

1. INTRODUCTION

Virtually all physical processes involve variability over space and time. This variability can be quite complicated, including variability caused by spatial nonstationarities, and nonseparable space-time interaction. Furthermore, the relevant datasets often include many spatial and temporal observations. Traditional approaches to this problem have focused on the geostatistical paradigm (Bilonick, 1983; Cressie & Majure, 1997), which requires complete specification of the joint space-time covariance structure, and on multivariate time-series methods, which specify dynamic models that are linked spatially (Rouhani & Wackernagel, 1990). The former approach is limited in that difficulties often arise in the specification and implementation of realistic covariance functions. The latter approach is limited in that it is difficult to predict at unmonitored sites because of the lack of a continuous spatial component in the model. If both temporal and spatial compo-
ments are present, it is natural to combine both approaches, in a statistical model that is temporally dynamic and spatially descriptive. We refer to such a model as a space-time dynamic model.

With the addition of a measurement error component in the model, the space-time dynamic processes can be described through a state-space formulation. It is then natural to consider prediction via a Kalman filter (Kalman, 1960). Generalising the Kalman filter to a spatial context has its difficulties. Space does not have a natural ordering and hence the dynamic updating that is so important to Kalman filtering is missing. Thus, Thiébaux’s (1991) suggestion that Kalman filtering in space follows analogously to that in time is not helpful. When there is also a temporal component, the spatial component can be used to index the vector of measurements and the vector of the state variable. This gives rise to a multivariate Kalman filter but of very high dimension, depending on the extent and resolution of the spatial component (Ghil et al., 1981). Our research addresses directly the problem of dimension reduction while also accounting for each of the components of variation that give rise to the spatio-temporal data.

Although Kalman filter approaches to space-time prediction have been considered by the scientific community for nearly two decades (Ghil et al., 1981), the statistics community only recently considered space-time processes from this perspective. Commenting on Handcock & Wallis’s (1994) Bayesian approach to geostatistical space-time modelling, Cressie (1994) suggested that a Kalman filter incorporating space and time would be a powerful way to apply the Bayesian paradigm to such processes. Similarly, Guttrop, Meiring & Sampson (1994) considering many of the notions of space-time dynamic modelling, but did not focus on a Kalman filter estimation approach, nor did they consider a separate measurement equation. A complete description of this approach can be found in Meiring, Guttrop & Sampson (1998). In a paper entitled ‘Challenges in multivariate spatio-temporal modeling’ in the Proceedings of the XVIIth International Biometric Conference, Hamilton, Ontario, Canada, C. Goodall and K. V. Mardia sketched the idea for a reduced dimension space-time Kalman filter, in which the state process is written in terms of some set of basis functions. This reduced dimensionality is critical for implementing the Kalman filter on very large space-time datasets. Huang & Cressie (1996) were perhaps the first to demonstrate successfully a full implementation of a space-time dynamic model with a Kalman filter, albeit with a relatively simple, i.e. separable, space-time structure. Cane et al. (1996) implemented a reduced dimension space-time dynamic model using empirical orthogonal function basis functions. Mardia et al. (1998) provide the details and full implementation of the general reduced dimension model suggested in their Biometric Conference paper referred to previously. They call this model the ‘kriged Kalman filter’. Recently, Wikle, Berliner & Cressie (1998) implemented a fully Bayesian space-time dynamic model.

The reduced dimension space-time Kalman filter approaches described above are especially valuable since they allow one to model spatially nonstationary and space-time nonseparable processes, and allow implementation on large datasets. However, for many physical/biological processes, it is often the case that there is significant small-scale spatial structure that does not have a discernible temporal evolution. The space-time Kalman filter approaches described above have no structural component that can describe such a process. The authors have either assumed that the component is zero or they have absorbed it into the measurement error component. In the case of the latter, the resulting predictions are then too smooth and the filters are suboptimal (Ruiz-Molina & Valderrama, 1996). In this paper, we consider explicitly not only the measurement error
and space-time dynamic components, but also a non-dynamic term that captures small-scale spatial variability. We then derive a general empirical Bayesian predictor based on this model, which has as special cases the space-time Kalman filter approaches described above and the simple kriging spatial prediction procedure (Cressie, 1993, p. 110). A more complete discussion of how this model compares to other recent space-time dynamic models can be found in Cressie & Wikle’s (1998) invited discussion of Mardia et al. (1998).

In § 2 we describe the temporally dynamic, spatially descriptive model, as well as the reduced-dimension representation and empirical Bayesian implementation. The estimation of model parameters is outlined in § 3. In § 4, the model is applied to a dataset of near-surface tropical winds.

2. Methodology

2.1. Preamble

Assume that data $Z(s_1; t_1), \ldots, Z(s_N; t_N)$ are obtained from an observable and spatially continuous process $Z(s; t)$, where $s \in D$, with $D$ some spatial domain in $d$-dimensional Euclidean space $R^d$ and $t \in \{1, 2, \ldots\}$, a discrete index of times. We suppose that the observable process has a component of measurement error expressed through the measurement equation

\[ Z(s; t) = Y(s; t) + \varepsilon(s; t), \]  

where $Y(s; t)$ can be thought of as an unobservable process, ‘smoother’ than $Z(s; t)$. Our goal is to predict the process $Y(\cdot; \cdot)$ at all spatial locations and time points of interest, regardless of where and when the data $Z(s_1; t_1), \ldots, Z(s_N; t_N)$ are observed. In the following development, we assume that the data and the processes we are dealing with have been suitably detrended.

2.2. A model for the unobservable space-time process

Assume that $Y(s; t)$ can be written

\[ Y(s; t) = Y_k(s; t) + \nu(s; t), \]  

where $\nu(s; t)$ is a component of variance representing small-scale spatial variation that does not have temporally dynamic structure, i.e. a ‘spatially descriptive’ component. By contrast, the component $Y_k(s; t)$ is assumed to evolve according to the state equation

\[ Y_k(s; t) = \int_D w_s(u)Y_k(u; t - 1)\,du + \eta(s; t), \]  

where $\eta(s; t)$ is a spatially coloured error process, a ‘spatially descriptive’ component, and $w_s(u)$ is a function representing the interaction between the state process $Y_k$ at location $u$ and time $(t - 1)$ and $Y_k$ at location $s$ and time $t$, a ‘temporally dynamic’ component. For stationarity over time, we further require that this interaction function satisfies $\int_D w_s(u)\,du = \pi$, where $|\pi| < 1$ is an unknown parameter. This decomposition of $Y$ into a space-time dynamic component $Y_k$ and a small-scale spatial component $\nu$ is important and, as discussed in § 1, has not been considered in other space-time Kalman filter methodologies.

Additionally, we assume that the $\varepsilon$ process is uncorrelated with $Y$, $\eta$ and $\nu$ over all locations and time lags. We also assume that $E\{\varepsilon(s; t)\varepsilon(r; \tau)\} = 0$, $E\{\nu(s; t)\nu(r; \tau)\} = 0$, and $E\{Y(s; t)\nu(r; \tau)\} = 0$. 
E\{η(s; t)η(r; τ)\} = 0, for all s, r, t, τ; E\{v(s; t)η(r; τ)\} = 0, for all r, s, t, τ; and
E\{v(s; t)Y_K(r; t)\} = 0, E\{η(s; t)Y_K(r; t - 1)\} = 0, for all r, s, t.

We assume that \( Y_K \) can be decomposed into \( K \) dominant components,

\[
Y_K(s; t) \equiv \sum_{k=1}^K \phi_k(s)a_k(t),
\]

for \( s \in D \) and \( t \in \{1, 2, \ldots\} \), where \( a_k(\cdot) \) \((k = 1, \ldots, K)\) are zero-mean time series, and \( \{\phi_i(\cdot); i = 1, 2, \ldots\} \) are deterministic basis functions that are complete and orthonormal, that is \( \int_D \phi_i(u)\phi_j(u) \, du = \delta_{ij} \), where \( \delta_{ij} = 0 \) for \( i \neq j \) and \( \delta_{ij} = 1 \) for \( i = j \). Furthermore, we take advantage of the completeness of the \( \phi \)'s and expand the interaction function as

\[
w_i(u) = \sum_{i=1}^\infty b_i(s)\phi_i(u),
\]

for \( s, u \in D \) and where \( \{b_i(s); l = 1, 2, \ldots\} \) are unknown but nonstochastic parameters.

Substituting (5) and (4) into (3) and making use of the orthonormality property, we obtain

\[
Y_K(s; t) = b(s)'a(t - 1) + \eta(s; t),
\]

where \( b(s) \equiv (b_1(s), \ldots, b_K(s))' \) and \( a(t - 1) \equiv (a_1(t - 1), \ldots, a_K(t - 1))' \). Noting that the temporally dynamic component (4) can also be written \( Y_K(s; t) = \phi(s)'a(t) \), where \( \phi(s) \equiv (\phi_1(s), \ldots, \phi_K(s))' \), we substitute this into (6) to obtain

\[
\phi(s)'a(t) = b(s)'a(t - 1) + \eta(s; t).
\]

We can also rewrite the measurement equation (1) as

\[
Z(s; t) = \phi(s)'a(t) + v(s; t) + e(s; t).
\]

As a consequence of the original model assumptions, we have \( E\{a_i(t)v(s; \tau)\} = 0 \), for all \( i = 1, \ldots, K, s, t, \tau \), and \( E\{a_i(t)e(s; \tau)\} = 0 \), for all \( i = 1, \ldots, K, s, t, \tau \).

Assume that we have \( N = n \times T \) data at locations \( \{s_1, \ldots, s_n\} \) and at time points \( \{1, \ldots, T\} \), as one might obtain from a network of monitoring stations. In fact, our approach is flexible enough to handle missing data or mobile monitoring sites. For the case considered here, we can write (7) as a system of \( n \) equations, one for each \( s_i \), for \( i = 1, \ldots, n \). In matrix form, this linear system is written as

\[
\Phi a(t) = Ba(t - 1) + \eta(t),
\]

where \( \eta(t) \equiv (\eta(s_1; t), \ldots, \eta(s_n; t))' \), and where we define the \( n \times K \) matrices \( \Phi \equiv \{\phi(s_1), \ldots, \phi(s_n)\}' \) and \( B \equiv \{b(s_1), \ldots, b(s_n)\}' \). Assuming \( n \geq K \) and \((\Phi'\Phi)^{-1}\) nonsingular, we can write (9) as

\[
a(t) = Ha(t - 1) + J\eta(t),
\]

where \( J \equiv (\Phi'\Phi)^{-1}\Phi' \) and the \( K \times K \) matrix \( H \equiv JB \). Note that, if the truncation parameter \( K \) is allowed to vary with time, then \( H \) becomes time-varying.

2-3. The optimal predictor

Recall that \( Y_K(s; t) = \phi(s)'a(t) \) and hence, if an optimal, i.e. minimum mean-squared prediction error, predictor of \( a(t) \) is found, the optimal predictor of \( Y_K(s; t) \) is immediately available after premultiplying by \( \phi(s)' \). Then the optimal predictor for \( Y(s; t) \) follows in a
straightforward manner as we shall see below. This optimal predictor is linear in the Gaussian case. Furthermore, if the Gaussian assumption is not appropriate, this linear predictor is optimal in the class of all linear predictors.

Given the measurement equation (8) and the state-process equation (10), the optimal predictor of \( a(t) \) given observations up to and including \( t \) is expressed recursively in terms of a Kalman filter (Kalman, 1960; Meinhold & Singpurwalla, 1983):

\[
\hat{a}(t \mid t) = E\{a(t) \mid Z(t), \ldots, Z(1)\} = \hat{a}(t \mid t-1) + G(t)\{Z(t) - \Phi \hat{a}(t \mid t-1)\},
\]

for \( t \geq 1 \), with mean-squared prediction error

\[
P(t \mid t) = E[\{a(t) - \hat{a}(t \mid t)\}{a(t) - \hat{a}(t \mid t)}'] = P(t \mid t-1) - G(t)\Phi P(t \mid t-1),
\]

where \( Z(t) \equiv (Z(s_1; t), \ldots, Z(s_n; t))' \) and the Kalman gain \( G(t) \) is given by

\[
G(t) = P(t \mid t-1)\Phi'\{R + V + \Phi P(t \mid t-1)\Phi\}^{-1}.
\]

The one-step-ahead predictions are given by

\[
\hat{a}(t \mid t-1) = E\{a(t) \mid Z(t-1), \ldots, Z(1)\} = H\hat{a}(t-1 \mid t-1),
\]

\[
P(t \mid t-1) = \text{var}\{a(t) \mid Z(t-1), \ldots, Z(1)\} = HP(t-1 \mid t-1)H' + JQJ',
\]

where \( R = \text{var}\{\varepsilon(t)\} \), \( V = \text{var}\{\nu(t)\} \), \( Q = \text{var}\{\eta(t)\} \), and where

\[
\varepsilon(t) \equiv (\varepsilon(s_1; t), \ldots, \varepsilon(s_n; t))', \quad \nu(t) \equiv (\nu(s_1; t), \ldots, \nu(s_n; t))',
\]

and \( \eta(t) \) was similarly defined earlier. To start the Kalman recursion, we assume that \( \hat{a}(0 \mid 0) \equiv 0 \) and \( P(0 \mid 0) \equiv \hat{C}_0^{\hat{\kappa}} \), where in practice an estimate \( \hat{C}_0^{\hat{\kappa}} \), given in § 3·4, is substituted for \( C_0^{\hat{\kappa}} \).

Now, consider prediction of the process \( Y(s; t) \) based on the Kalman filter predictor \( \hat{a}(t \mid t) \). The optimal predictor is then

\[
\hat{Y}(s; t \mid t) = \phi(s)\hat{a}(t \mid t) + c_v(s)(C_v^0)^{-1}Z(t),
\]

where

\[
C_v^0 \equiv \text{cov}\{Z(t), Z(t)\}, \quad c_v(s) \equiv E\{\nu(s; t)\nu(t)\} = (c_v(s, s_1), \ldots, c_v(s, s_n))'
\]

and \( c_v(s, r) \equiv E\{\nu(s; t)\nu(r; t)\} \). The proof of this result is given in the Appendix. We note that the second term in (16) is a type of simple kriging (Cressie, 1993, p. 110) applied to the spatial error term \( \nu(s; t) \). Thus, as the truncation integer \( K \) decreases, the optimal predictor of \( Y(s; t) \) begins to look more and more like the simple-kriging predictor in the presence of measurement error. This formula is to be contrasted with that of C. Goodall and K. V. Mardia, in their conference paper mentioned in § 1, and Mardia et al. (1998), which does not have the second term. The recognition of the component of variation \( \nu \) in (2) and its presence in the optimal predictor (16) is important.

The conditional prediction error variance for \( Y(s; t) \) is

\[
\text{var}\{Y(s; t) - \hat{Y}(s; t \mid t) \mid Z(t), \ldots, Z(1)\} = \phi^2(s)P(t \mid t)\phi(s) + c_v(s, s) - c_v(s)'(C_v^0)^{-1}c_v(s)
\]

\[
- 2\phi(s)'\text{cov}\{\hat{a}(t \mid t), Z(t)\}(C_v^0)^{-1}c_v(s),
\]

where the proof of this result is given in the Appendix. The first term in (17) corresponds to the prediction error variance from the \( Y_k \) process; the second and third terms together correspond to the simple-kriging prediction variance of the \( \nu \) process; the last term is a correction resulting from the covariance between the Kalman filter prediction of the \( Y_k \)
process, through the Kalman filter prediction of $a$, and the simple-kriging predictor. As the truncation integer $K$ decreases, the prediction error variance tends to behave more and more like the simple-kriging variance.

The one-step-ahead optimal predictor and prediction error variance for $Y(s; \tau + 1)$ can be found in a similar fashion. In this case, the kriging terms do not contribute to the one-step-ahead prediction since the $v$ process is assumed not to evolve in time.

3. **Estimation of model parameters**

3-1. **Introduction**

The predictor presented in § 2.3 is optimal assuming we know the covariances $R$, $V$ and $Q$, as well as the state matrix $H$, that is $\Phi$ and $B$. In our case, we can choose the $\phi$'s, see § 3.5, as long as they are complete and orthonormal, but we must estimate the $R$, $V$, $Q$ and $B$ matrices. By substituting these estimates in place of the unknown matrices, we no longer obtain exactly the conditional expectation. This approach is completely analogous to the practice of Kalman filtering in time or kriging in space, where the covariance or variogram parameters are usually unknown and have to be estimated. A general discussion of the problems in obtaining predictive distributions given estimated parameters can be found in Zimmerman & Cressie (1992). Our approach corresponds to viewing the Kalman filter as an empirical Bayesian technique (Meinhold & Singpurwalla, 1983). Alternatively, fully Bayesian hierarchical approaches to space-time dynamic modelling can be implemented (Wikle et al., 1998), but there is a trade-off between computational efficiency, with the empirical Bayesian approach, and statistical precision, with the fully Bayesian approach. Our approach is motivated by the need to model large spatio-temporal datasets. Hence, we focus on computationally efficient empirical Bayesian procedures.

We use standard method of moments estimators for $R$, $V$, $Q$ and $B$. Although maximum likelihood estimators of model parameters are more efficient, the high-dimensional nature of spatio-temporal problems makes for poorly behaved likelihood surfaces and iterative algorithms that are difficult to implement, thus defeating one of our goals of being able to process efficiently large spatio-temporal datasets.

3-2. **Estimation of model covariances**

In § 2, the space-time model is developed under the assumption that the measurement error covariance matrix $R$ and the residual spatial process covariance matrix $V$ can take the form of any valid covariance matrix. For the purposes of estimation, we now assume that the dynamic space-time component of the model, $Y_K$, contains any nonstationary spatial structure, through $H$ and $\eta(t)$. Then we assume that the non-dynamic spatial component will be composed of a nugget effect, which here is the variance of the $e$'s, and a stationary isotropic process, $v$. Thus, we let $R = \sigma_v^2 I$ and estimate $\sigma_v^2$ preferably from information about the measuring instrument, or by the behaviour of an empirical variogram estimate of the data as the spatial lag approaches zero; for a discussion of variograms, their definition, estimation and modelling, see Cressie (1993, § 2.4).

Given $C_0^Y \equiv \text{cov} \{Y(t), Y(t)\}$ and $C_0^X \equiv \text{cov} \{Y_k(t), Y_k(t)\} = \Phi J C_0^X J' \Phi'$, the residual covariance matrix associated with the $v$ process is given by $V = C_0^Y - C_0^X$. In practice, given an estimate of $C_0^Y$ and the specified basis functions $\Phi$, we can obtain an empirical estimate of $V$, $\hat{V}_E \equiv \hat{C}_0^Y - \hat{C}_0^X$, where $\hat{C}_0^X = \Phi J \hat{C}_0^Y J' \Phi'$, and $\hat{C}_0^Y$ is obtained as in § 3-4. As is well known from the estimation of variance component models by the method of moments, one must take care to ensure that $\hat{V}_E$ is positive definite. In practice, this is achieved by
examining different choices for the measurement error variance and/or the truncation parameter $K$ in (4). The matrix $\hat{V}_k$ only gives estimates of the $v$-process covariances at locations for which we have data. We use this empirical estimate to fit a valid, i.e. positive-definite, spatially stationary and isotropic model, $c_v(s, r) \equiv \text{cov}\{v(s; t), v(r; t)\}$. Should the empirical estimate $\hat{V}_k$ suggest a diagonal, or possibly diagonally dominant, matrix we can assume that $v$ is simply white noise, and so $c_v(s, r) = \sigma_v^2$, for $s = r$, and zero otherwise. Identifiability for the $v$-process comes from the declaration of the number of components $K$ in (4). In this case, the $v$-process is, apart from measurement error, everything that we are not willing to declare is dynamic in nature. In practice, when one does not have much experience with a given type of data, the choice of $\sigma_v^2$ and $K$, and hence $\sigma_v^2$, proceeds in an exploratory manner, much as one would choose dominant components in a principal components analysis.

Examination of the Kalman filter equations (11)–(15) shows that we need only estimate $JQJ'$, rather than $Q$. From (10), it can be shown that

$$JQJ' = J(C_0^Z - V - R)J' - JC_1^Z J'H' - HJ(C_1^Z)'J' + HJ(C_0^Z - V - R)J'H',$nolabel

(18)

where $C_0^Z \equiv \text{cov}\{Z(t), Z(t)\}$ and $C_1^Z \equiv \text{cov}\{Z(t), Z(t-1)\}$. We have assumed temporal invariance; i.e. lag-zero and lag-one temporal covariances do not depend on $t$. To obtain the estimate $\hat{J} \hat{Q} \hat{J}'$, we can substitute estimates for $C_0^Z$, $C_1^Z$, $V$, $R$ and $H$ into (18), where once again care is taken to ensure positive definiteness of $(\hat{C}_0^Z - \hat{V} - \hat{R})$ and $\hat{J} \hat{Q} \hat{J}'$. Estimation of $H = JB$, $C_0^Z$ and $C_1^Z$ is discussed below.

### 3-3. Estimation of the state parameter matrix

From (2), (6) and (9), we can write $Y(t) = Ba(t - 1) + \eta(t) + v(t)$, for $t \geq 1$. Post-multiplying this by $Y(t - 1)'$ and taking expectations, we obtain $C_Y = BJ(C_0^Z - V - R)J' \Phi'$, where we have exploited the independence relationships of § 2-2 and used

$$C_Y = E\{Y(t)Y(t-1)\}.$$

Noting that $C_1^Y = C_Y'$, we can write

$$B = C_1^Y J'(J(C_0^Z - V - R)J')^{-1}.$nolabel

(19)

We then substitute estimates $\hat{C}_0^Z$, $\hat{C}_1^Z$, $\hat{V}$ and $\hat{R}$ into (19) to obtain

$$\hat{B} = \hat{C}_1^Z J'(J(\hat{C}_0^Z - \hat{V} - \hat{R})J')^{-1},$$

with care being taken to ensure that $J(\hat{C}_0^Z - \hat{V} - \hat{R})J'$ is positive definite. Finally, $\hat{H} = J\hat{B}$.

### 3-4. Estimation of data and process covariances

As shown in §§ 3-2 and 3-3, to obtain estimates of the Kalman filter model parameters, we need estimates of the covariance matrices $C_0^Z$, $C_1^Z$ and $C_0^Y$. We estimate $C_0^Z$ and $C_1^Z$ using standard moment estimators. Then, from the estimate of $R$ given in § 3-2, we obtain the covariance matrix estimate of the $Y$ process as $\hat{C}_0^Y = \hat{C}_0^Z - \hat{R}$, with care being taken to ensure that $\hat{C}_0^Y$ is positive definite.

### 3-5. Basis functions

As stated previously, we can choose any set of basis functions $\{\phi_k(\cdot)\}$ in (4), as long as they are complete, orthonormal and are defined at any location $s$ in domain $D$. Thus,
there are many possible choices for these functions, such as orthogonal polynomials and wavelets. In this study, we consider an empirical orthogonal function basis set. These functions have a long history of being considered for spatial prediction in statistics (Cohen & Jones, 1969). Furthermore, Cretin & Obled (1982) point out that the empirical orthogonal function approach to spatial prediction naturally accounts for anisotropic and heterogeneous covariance structure. Using empirical orthogonal functions in the Kalman filter context was suggested by C. Goodall and K. V. Mardia, in their conference paper referred to in § 1, and implemented in Cane et al. (1996).

As discussed by Buell (1972), when data are evenly distributed in space, i.e. gridded, the empirical orthogonal function basis set is essentially equivalent to principal component vectors as defined in multivariate statistics. Our observation and state processes are continuous in space, and data are not in general evenly distributed spatially. However, the empirical orthogonal function basis functions can be easily obtained by a 'pre-gridding' procedure (Karl, Koscielny & Diaz, 1982). That is, we use some simple space-time prediction scheme to obtain smooth predictions of the Y process on the prediction grid of interest given the irregularly spaced data. Then we calculate the empirical orthogonal function basis functions from this field defined on the regular grid by performing a principal component decomposition. For example, let the gridded process be denoted by $Y^o(r_i; t)$ for $i = 1, \ldots, m$ and $t = 1, \ldots, T$, where $\{r_i : i = 1, \ldots, m\}$ are the regular grid locations at which basis functions must be defined. Write $Y^o(t) \equiv (Y^o(r_1; t), \ldots, Y^o(r_m; t))'$. Then the empirical orthogonal function basis functions can be obtained from the symmetric decomposition $\hat{C}_0 = \Psi^o \Lambda^o (\Psi^o)'$, where $\hat{C}_0$ is an estimate of the covariance matrix $C_0$ $\equiv E(Y^o(t)Y^o(t)')$ and $Y^o$ is assumed to have zero mean. The basis functions then correspond to the columns of the eigenmatrix $\Psi^o \equiv \{\psi_1^o, \ldots, \psi_m^o\}$, where $\psi_i^o \equiv (\psi_{i1}^o(r_1), \ldots, \psi_{im}^o(r_m))'$.

In practice, one can use any stochastic or nonstochastic prediction procedure to obtain $Y^o$. We consider a nonstochastic procedure to predict on a regular grid at each time and then smooth over time at each spatial location. Algorithmically, we first predict the Y process at each spatial location on the prediction grid for each data time $t = 1, \ldots, T$ using biharmonic splines (Sandwell, 1987). Next, for each spatial location $r_1, \ldots, r_m$ on the prediction grid, we smooth over time using a local Gaussian kernel smooth with the time-smoothing parameter chosen by generalised crossvalidation (Hastie & Tibshirani, 1990, pp. 49–52). This procedure gives $Y^o(t)$ $(t = 1, \ldots, T)$, from which the covariance estimate $\hat{C}_0$ is obtained via the method of moments. One can imagine more sophisticated space-time smoothing procedures. Various alternative approaches were considered and, for the application described in § 4, the resulting empirical orthogonal functions were found to be relatively insensitive to the choice of the space-time smoothing algorithm. This will not be the case in general. Alternatively, one could use the Karhunen–Loève expansion in continuous space to obtain the basis functions in the space-time Kalman filter context, as described in C. K. Wikle’s 1996 Ph.D. Dissertation from Iowa State University.

4. Near-surface wind example

4.1. Background to the data

To illustrate the implementation of the modelling methodology presented in §§ 2 and 3, we extracted wind data, valid for a height of 10 metres above the surface, from the National Centers for Environmental Prediction reanalysis product (Kalnay et al., 1996). While the subset of the data that we analyse is not huge, it is quite large, involving about 14 000
observations; our purpose in this section is more to determine the effectiveness of our spatio-temporal prediction method than to demonstrate its performance on huge datasets.

The dataset consists of the east-west component of the wind velocity vector from a region over the tropical western Pacific ocean (145°E–175°E, 14°S–16°N) for the period November 1992–February 1993. The winds are given every 6 hours and at a spatial resolution of nearly 2 degrees in latitude and longitude, or about 210 km by 210 km. This gives data on a $17 \times 17$ grid, that is 289 locations, and at 480 time points. To test our model, we selected 40 of the 289 grid locations at random, with the exception that no observation location is allowed within a $12^\circ$ by $12^\circ$ square centred in the middle of the grid. Our purpose in this study is to determine how well we predict the east–west wind speed component on the full $17 \times 17$ grid, using only the wind information at these 40 sites and 480 time points. The prediction grid and ‘observation’ locations are shown in Fig. 1.

![Fig. 1. Prediction grid (+) and observation locations (*) for the tropical wind component dataset. The region is to the north-east of Australia and Papua New Guinea.](image)

Although the east–west surface winds over the tropics are not expected to exhibit a significant spatial trend, we removed the time-averaged mean from each grid location to create a zero-mean dataset from which to evaluate our model. Furthermore, to test the model’s ability to perform under various signal-to-noise-ratio, \( \text{SNR} \), conditions, where \( \text{SNR} \) is the variance of the \( Y \) process divided by the measurement error variance, we created two synthetic observation datasets by adding white noise with different variances to the original measurements. Specifically, we considered \( \text{SNR} = 10 \) and 5, which correspond to measurement error variances of 0.7 and 1.4 (m/sec)$^2$, respectively.

4-2. **Descriptive results of model prediction**

In this demonstration, the measurement error variance is known. In practice, one must estimate this as described in § 3-2. The truncation parameter was chosen to be $K = 20$;
based on an examination of the eigenvalues of the estimate $\hat{C}_0^{\nu}$ given in § 3.4, this value of $K$ accounts for approximately 94% of the variance in the $Y$ process at observation locations, yet still represents a large reduction in the dimensionality, from 289 down to 20, of the state process. Estimates of the $\nu$-process covariance matrix $\tilde{V}$ were then found and a model was fitted. Figure 2 shows the estimated $\nu$ covariances as a function of lag distance, the bin averages of those estimated covariances and the fitted covariance function. Note the negative covariances in the 500–1000 km spatial-lag region. This structure is common in meteorological variables (Daley, 1991, § 4.3). Since the known class of valid negative-dependence covariance models is small, and to facilitate automation of our methodology, we fitted a nonparametric isotropic covariance function, as described in Shapiro & Botha (1991).

![Figure 2](image)

Fig 2. Empirically estimated covariances (points) of the $\nu$ process given in (2), with truncation parameter $K = 20$ and $\text{SNR} = 5$; bin averages of covariance estimates are shown by circles, and fitted nonparametric covariance model by the solid line.

Figure 3(a) shows a spatial cross-section over time for the low SNR noisy data, Fig. 3(b) the predicted signal based on the noisy data, Fig. 3(c) the true signal and Fig. 3(d) the true signal minus the predicted signal. The cross-section was taken at 8.5°N latitude. The model has done a credible job of capturing the essence of the spatio-temporal signal from relatively few noisy observations. Of course, there are differences between the predicted and the true signals, generally reflecting the fact that the predicted signal tends to smooth extreme events, as is typically the case with Kalman filter predictions.

Figure 4 shows the time-average model-predicted root mean squared prediction error for the case of $\text{SNR} = 5$. The largest values are found in the centre ‘hole’ region and along the boundary, as expected. Note that these prediction errors do not exhibit the ‘bull’s-eye’ character typical of kriging predictors. This is due to the large-scale nature of the leading empirical orthogonal function basis functions, which tends to make the prediction errors more uniform.
Fig. 3. Shown are $x$–$t$ image plots at latitude $8.5^\circ$N for (a) the noisy process $Z$ (SNR = 5), (b) prediction $\hat{Y}$ of the true signal based on the noisy observations, (c) the true signal $Y$ and (d) the truth minus the prediction ($Y - \hat{Y}$). Darker shading is indicative of negative component wind speeds (m/sec) and lighter shading is indicative of positive component wind speeds; all plots use identical grey scales.

Fig. 4. The model-estimated root mean square prediction errors for the $Y$-process in (2), based on data from the $Z$-process with SNR = 5. The contour interval is 0.2 m/sec. The observation locations are indicated by *.
4-3. Objective results of model prediction

For a quantitative measure of a model's precision and accuracy that can facilitate comparison, we consider the following three validation statistics (Carroll & Cressie, 1996):

\[
CR_1(s_j) = \frac{(1/T) \sum_{t=1}^{T} \{Y(s_j; t) - \hat{Y}(s_j; t)\}}{(1/T) \sum_{t=1}^{T} \delta Y(s_j; t)\frac{2}{3}},
\]

\[
CR_2(s_j) = \left( \frac{(1/T) \sum_{t=1}^{T} \{Y(s_j; t) - \hat{Y}(s_j; t)\}^2}{(1/T) \sum_{t=1}^{T} \delta Y(s_j; t)} \right)^{\frac{3}{2}},
\]

\[
CR_3(s_j) = \left( \frac{1}{(1/T) \sum_{t=1}^{T} \{Y(s_j; t) - \hat{Y}(s_j; t)\}^2} \right)^{\frac{3}{2}},
\]

where \(\hat{Y}(s_j; t)\) is the prediction of the process \(Y\) at location \(s_j\) and time \(t\), and where \(s_j\) \((j = 1, \ldots, m)\) are the \(m = (17)^2\) prediction locations on the regular grid. Furthermore, \(\delta Y(s_j; t)\) is the corresponding root mean squared prediction error. The statistic \(CR_1(s_j)\) provides an estimate of the unbiasedness of the predictors for each prediction location and should be close to 0; \(CR_2(s_j)\) gives a measure of the accuracy of the mean squared prediction errors and should be close to 1; and \(CR_3(s_j)\) is a measure of the 'goodness of prediction'. One would like \(CR_3(s_j)\) to be small so that the predicted values are close to the true values. One could then look at spatial plots of (20), (21) and (22) as a function of location \(s_j\) or one could look at spatial averages. We will take the latter approach in this presentation.

Table 1 shows the validation statistics for the two values of SNR. In addition, results are given for the case when \(v\) is modelled using nonparametric covariances and when it is assumed to be spatial white noise. Note that the 'ave' of the \(CR\) statistics denotes the spatial average. As is evident from the ave\(\langle CR_1 \rangle\) statistic, each implementation produces relatively unbiased predictions. For comparison of a model's ability to predict the true wind, ave\(\langle CR_3 \rangle\) is the relevant statistic. On the basis of such a comparison, the percentage improvement is larger as SNR decreases. In addition, the model performs slightly better, by 2–4%, when \(v\) is modelled using nontrivial covariances as compared to when \(v\) is assumed to be spatial white noise.

<table>
<thead>
<tr>
<th>Method</th>
<th>SNR</th>
<th>ave(\langle CR_1 \rangle)</th>
<th>ave(\langle CR_2 \rangle)</th>
<th>ave(\langle CR_3 \rangle)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model (v)</td>
<td>5</td>
<td>0-0023</td>
<td>1-4308</td>
<td>1-2563</td>
</tr>
<tr>
<td>White (v)</td>
<td>5</td>
<td>0-0025</td>
<td>1-3880</td>
<td>1-2845</td>
</tr>
<tr>
<td>Model (v)</td>
<td>10</td>
<td>0-0014</td>
<td>1-4596</td>
<td>1-6464</td>
</tr>
<tr>
<td>White (v)</td>
<td>10</td>
<td>0-0018</td>
<td>1-4037</td>
<td>1-2107</td>
</tr>
</tbody>
</table>

The comparison between modelled and empirical root mean squared prediction error can be examined via ave\(\langle CR_2 \rangle\) as shown in Table 1. The model tends to underestimate this quantity, probably because of the various parameters fitted; see § 3. It is interesting to note that comparisons with predictions made by applying simple kriging to each time period independently have shown that the new model performs better, in terms of a lower \(CR_3\) statistic, for these data; see C. K. Wikle and N. A. Cressie's unpublished technical report 97-24 in the Statistical Laboratory Series at Iowa State University, Ames, Iowa,
USA. Although such comparisons are dataset-dependent, validation statistics such as (20)–(22) provide convenient and objective tools for comparison.

ACKNOWLEDGEMENT

Wikle’s research was sponsored by the U.S. Department of Energy, Office of Energy Research, Environmental Sciences Division, Office of Health and Environmental Research, under his appointment to the Graduate Fellowships for Global Change administered by Oak Ridge Institute for Science and Education. Further support was provided by the National Center for Atmospheric Research Geophysical Statistics Project, sponsored by the National Science Foundation. Cressie’s research was supported in part by a co-operative agreement between the U.S. Environmental Protection Agency and Iowa State University.

APPENDIX

Proof of (16) and (17)

Equation (16) is derived as follows:

\[ \hat{Y}(s; t|t) \equiv E\{Y(s; t)|Z(t), Z^*(t-1)\} \]
\[ = E\{\phi(s)'a(t) + v(s; t)|Z(t), Z^*(t-1)\} \]
\[ = \phi(s)'E\{a(t)|Z(t), Z^*(t-1)\} + E\{v(s; t)|Z(t), Z^*(t-1)\} \]
\[ = \phi(s)'a(t|t) + c_v(s)(C_v^Z)^{-1}Z(t), \tag{A1} \]

where

\[ Z^*(t-1) \equiv \{Z(t-1), \ldots, Z(1)\}, \quad C_v^Z \equiv \text{cov} \{Z(t), Z(t)\}, \]
\[ c_v(s) \equiv E\{v(s; t)v(t)\} = \{c_v(s, s_1), \ldots, c_v(s, s_n)\}', \quad c_v(s, r) \equiv E\{v(s; t)v(r; t)\}. \]

The relationship \( E\{v(s; t)|Z(t), Z^*(t-1)\} = c_v(s)(C_v^Z)^{-1}Z(t) \) is easily seen if \( v(s; t) \) and \( Z(t) \) have a joint multivariate Gaussian distribution and we note that \( E\{v(s; t)Z(r; t)\} = c_v(s, r) \).

The conditional prediction error variance for \( \hat{Y}(s; t) \) is given by

\[ \text{var} \{Y(s; t) - \hat{Y}(s; t|t)|Z(t), Z^*(t-1)\} = \text{var} \{Y(s; t)|Z(t), Z^*(t-1)\}. \]

Then, from (2) and (4),

\[ \text{var} \{Y(s; t)|Z(t), Z^*(t-1)\} = \text{var} \{\phi(s)'a(t) + v(s; t)|Z(t), Z^*(t-1)\} \]
\[ = \text{var} \{\phi(s)'a(t)|Z(t), Z^*(t-1)\} + \text{var} \{v(s; t)|Z(t), Z^*(t-1)\} \]
\[ + 2 \text{cov} \{\phi(s)'a(t), v(s; t)|Z(t), Z^*(t-1)\}. \tag{A2} \]

Now, from (12), note that

\[ \text{var} \{\phi(s)'a(t)|Z(t), Z^*(t-1)\} = \phi(s)'P(t|t)\phi(s), \tag{A3} \]

and, following the assumptions used to get (A1), we obtain

\[ \text{var} \{v(s; t)|Z(t), Z^*(t-1)\} = c_v(s, s) - c_v(s)(C_v^Z)^{-1}c_v(s). \tag{A4} \]

Regarding the third term, note that \( \text{cov} \{\phi(s)'a(t), v(s; t)\} = 0 \) by the assumptions of § 2.2 and

\[ \text{cov} \{\phi(s)'a(t), v(s; t)\} = E[\text{cov} \{\phi(s)'a(t), v(s; t)|Z(t), Z^*(t-1)\}] \]
\[ + \text{cov} [E\{\phi(s)'a(t)|Z(t), Z^*(t-1)\}]. \tag{A5} \]
Since
\[ E[\text{cov}\{\phi(s')a(t), v(s; t)|Z(t), Z^*(t-1)\}] = \text{cov}\{\phi(s')a(t), v(s; t)|Z(t), Z^*(t-1)\}, \]
because of the Gaussianity assumption, we can write
\[
\begin{align*}
\text{cov}\{\phi(s')a(t), v(s; t)|Z(t), Z^*(t-1)\} &= -\text{cov}[E\{\phi(s')a(t)|Z(t), Z^*(t-1)\}, E\{v(s; t)|Z(t), Z^*(t-1)\}] \\
&= -\text{cov}\{\phi(s')a(t|t), c_\epsilon(s')(C_\epsilon^2)^{-1}Z(t)\} \\
&= -\phi(s')\text{cov}\{a(t|t), Z(t)\}(C_\epsilon^2)^{-1}c_\epsilon(s). \quad \text{(A6)}
\end{align*}
\]
Finally, substituting (A3), (A4) and (A6) into (A2), we arrive at
\[
\text{var}\{Y(s; t) - \hat{Y}(s; t|t)|Z(t), \ldots, Z(1)\} = \phi(s')P(t|t)\phi(s) + c_\epsilon(s, s) - c_\epsilon(s')(C_\epsilon^2)^{-1}c_\epsilon(s) - 2\phi(s')\text{cov}\{a(t|t), Z(t)\}(C_\epsilon^2)^{-1}c_\epsilon(s),
\]
which is (17).

\textbf{References}


[Received November 1997. Revised June 1999]