Environmental and Ecological Statistics (2020) 27:631–648
https://doi.org/10.1007/s10651-020-00447-3

Improved prediction for a spatio-temporal model

Gen Nowak[1] · A. H. Welsh[1]

Received: 20 October 2019 / Revised: 11 March 2020 / Published online: 30 June 2020

Abstract

We investigate a framework for improving predictions from models for spatio-temporal data. The framework is based on minimising the mean squared prediction error and can be applied to many models. We applied the framework to a model for monthly rainfall data in the Murray-Darling Basin in Australia. Across a range of prediction situations, we improved the predictive accuracy compared to predictions using only the expectation given by the model. Further, we showed that these improvements in predictive accuracy were maintained even when using a reduced subset of the data for generating predictions.

Keywords Kriging · Mixed model · Prediction · Spatio-temporal data

1 Introduction

Spatio-temporal data is becoming increasingly prevalent in areas ranging from climatology, ecology, epidemiology, cellular biology to the social sciences. As such, developing models for spatio-temporal data that are able to describe the structures present in the data and produce accurate predictions is an important area of statistical research. In this paper, we apply a framework for improving predictions for models for spatio-temporal data to modelling rainfall data in the Murray-Darling Basin (MDB) in Australia. However, the framework can be applied generally to models for spatio-temporal data.

Before detailing the framework for improving predictions, we first provide a brief outline of the general approach used for modelling spatio-temporal data. Let $Y(\boldsymbol{s}_i, t)$ denote the spatio-temporal data value at spatial location $\boldsymbol{s}_i$ and time point t, where $i = 1, \ldots, N$ and $t = 1, \ldots, T$. In matrix notation, we can express the data as a $T \times N$

Handling Editor: Bryan F. J. Manly.

✉ Gen Nowak
gen.nowak@anu.edu.au

[1] Research School of Finance, Actuarial Studies and Statistics, ANU College of Business and Economics, The Australian National University, Acton, ACT 2601, Australia

matrix $\boldsymbol{Y}$, where the rows correspond to time points and the columns correspond to spatial locations. As a general model for $Y(\boldsymbol{s}_i, t)$, following Banerjee et al. (2015), we write

$$Y(\boldsymbol{s}_i, t) = \mu(\boldsymbol{s}_i, t) + e(\boldsymbol{s}_i, t), \tag{1}$$

where $\mu(\boldsymbol{s}_i, t) = E(Y(\boldsymbol{s}_i, t))$ denotes the mean spatio-temporal structure and $e(\boldsymbol{s}_i, t)$ denotes the zero-mean residual component. Additional models and distributional assumptions can be applied to both the mean structure $\mu(\boldsymbol{s}_i, t)$ and the residual component $e(\boldsymbol{s}_i, t)$.

Consider now the problem of predicting a new data value $Y^* = Y(\boldsymbol{s}^*, t^*)$, at some spatial location $\boldsymbol{s}^*$ and time point t^*, given some training data $\boldsymbol{Y}^{\mathrm{tr}}$. A common approach is to use the estimated expected value $\hat{\mu}(\boldsymbol{s}^*, t^*) = \hat{E}(Y^*)$ as the prediction for Y^*. Depending on how $\mu(\boldsymbol{s}_i, t)$ in (1) was further modelled, computing the expected value $\hat{\mu}(\boldsymbol{s}^*, t^*)$ may require estimating various model parameters from the training data. A better approach to prediction is based on minimising the mean squared prediction error. If $g(\boldsymbol{Y}^{\mathrm{tr}})$ denotes a predictor of Y^*, the mean squared prediction error (MSPE) is given by

$$E\left[\left(Y^* - g\left(\boldsymbol{Y}^{\mathrm{tr}}\right)\right)^2\right].$$

The predictor that minimises the MSPE is $E\left(Y^* | \boldsymbol{Y}^{\mathrm{tr}}\right)$, the conditional expectation of Y^* given the training data $\boldsymbol{Y}^{\mathrm{tr}}$. Under some distributional assumptions on the data, we can derive an expression for $E\left(Y^* | \boldsymbol{Y}^{\mathrm{tr}}\right)$. If Y^* and $\boldsymbol{y}^{\mathrm{tr}}$ are jointly normally distributed, where $\boldsymbol{y}^{\mathrm{tr}}$ denotes the vectorised form of the training data, it can be shown that

$$E\left(Y^* | \boldsymbol{y}^{\mathrm{tr}}\right) = E(Y^*) + \mathrm{Cov}(Y^*, \boldsymbol{y}^{\mathrm{tr}})\mathrm{Var}(\boldsymbol{y}^{\mathrm{tr}})^{-1}\left(\boldsymbol{y}^{\mathrm{tr}} - E(\boldsymbol{y}^{\mathrm{tr}})\right). \tag{2}$$

This is of course the familiar expression for Best Linear Unbiased Predictors (BLUPS; Henderson 1950; Robinson 1991) that also underlies the method of kriging for spatial interpolation (see Krige 1951, 1962; Cressie 1990). Its advantage over the estimated expected value (the first term in (2)) is that it brings in and optimally weights information from observations in the training data that are correlated with the new data value. Note that the normality assumption is not a strict requirement to use (2) for producing predictions as it can also be derived without assuming normality (Harville 1976). However in this case, the sense in which the predictions are "best" is weaker because it is constrained by the linearity and unbiasedness requirements. Nonetheless, provided the data are approximately normal (or at least not strongly non-normal), then predictions from (2) are likely to be an improvement over using only $E(Y^*)$. Further, these improved predictions can be produced for many models.

To implement (2), we need to compute estimates of the $E(Y^*)$, $\mathrm{Cov}(Y^*, \boldsymbol{y}^{\mathrm{tr}})$, $\mathrm{Var}(\boldsymbol{y}^{\mathrm{tr}})$ and $E(\boldsymbol{y}^{\mathrm{tr}})$ terms, which will require estimating any necessary model parameters from the training data. Note that this framework for improved predictions will apply to any model where we can compute these terms. Ideally, we would like to use all of the training data to compute these terms because the optimality property requires doing so. However, depending on the size of the training data, computing the inverse of the $NT \times NT$ matrix $\mathrm{Var}(\boldsymbol{y}^{\mathrm{tr}})$ can quickly become computationally prohibitive. We will consider using a reduced subset of the training data for computing estimates of $\mathrm{Cov}(Y^*, \boldsymbol{y}^{\mathrm{tr}})$, $\mathrm{Var}(\boldsymbol{y}^{\mathrm{tr}})$ and $E(\boldsymbol{y}^{\mathrm{tr}})$, as this will decrease the number of calculations required. For example, we could use only recent time points or only neighbouring spatial locations in the training data. Depending on the type of prediction that is desired (e.g., predicting only at an unobserved time point, predicting only at an unobserved spatial location, or predicting at both an unobserved time point and spatial location), a particular method of subsetting the training data may be more appropriate than others. We explore various subsetting methods and compare their resulting predictive performance.

The motivation for improving prediction for spatio-temporal data stemmed from our work in Nowak et al. (2018). There we developed a hierarchical model for spatio-temporal rainfall data of the type suggested by Banerjee et al. (2015) and produced predictions in time and predictions in space. While the model performed reasonably well when predicting in time, it was outperformed by a naive single nearest neighbour-based prediction when predicting in space. This indicated that we were likely not sufficiently using the structure/information present in the data and that there was substantial scope to improve the predictions.

We apply the framework for improved prediction to a model for modelling spatio-temporal rainfall data in the MDB (described in Nowak et al. 2018). The predictions produced from applying the framework greatly improve the overall accuracy compared to those produced from the model using only $E(Y^*)$. In the next section, we provide some background on the MDB rainfall data and the model used for modelling these data. In Sects. 3.1, 3.2 and 3.3 we then detail the application of the framework for improved prediction to this model, for predicting in time, predicting in space and predicting in both space and time, respectively.

2 Model for monthly rainfall data in the Murray-Darling Basin

The data analysed consisted of monthly rainfall data recorded across a network of weather stations in the MDB from which high-quality data were available. Specifically, $Y(\boldsymbol{s}_i, t)$ denotes the cube-root transformed monthly rainfall measurement, recorded across $N = 78$ weather stations over a period of $T = 986$ months (spanning from January 1923 to February 2005). In addition to the rainfall measurements, a number of spatial variables were available for each station: the Cartesian x and y coordinates, the elevation and the climatic region (top, middle, bottom) to which the station belonged. Further details can be found in Nowak et al. (2018).

The hierarchical model we used for the MDB data proposes models for the mean and residual components in (1). The mean is modelled as

$$\mu(\boldsymbol{s}_i, t) = \sum_{j=0}^{J} \beta_j(\boldsymbol{s}_i) f_j(t), \tag{3}$$

where the $\left\{f_j(t)\right\}_{j=0}^{J}$ are smooth basis functions that describe temporal patterns that may be present in the data and $\left\{\beta_j(\boldsymbol{s}_i)\right\}_{j=0}^{J}$ are spatially-varying coefficients that allow the temporal patterns to differ across locations. We included $J = 3$ basis functions. We set $f_0(t) = 1$ and then let $f_1(t)$ and $f_2(t)$ represent deterministic seasonal effects. The final basis function is empirically derived by applying a singular value decomposition (SVD) to the residuals after removing the effects of the deterministic basis functions from the data and setting $f_3(t)$ to be the first left singular vector of the SVD.

The spatially varying coefficients $\left\{\beta_j(\boldsymbol{s}_i)\right\}_{j=0}^{J}$ are further modelled by assuming the $\boldsymbol{\beta}_j = \left(\beta_j(\boldsymbol{s}_1), \ldots, \beta_j(\boldsymbol{s}_N)\right)^T$ satisfy

$$\boldsymbol{\beta}_j \overset{\text{ind}}{\sim} N\left(\boldsymbol{X}_j \boldsymbol{\alpha}_j, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\boldsymbol{\beta}_j}}\right), \quad j = 0, \ldots, J, \tag{4}$$

where the columns of $\boldsymbol{X}_j$ represent a set of known spatial variables with $\boldsymbol{\alpha}_j$ denoting the unknown coefficients. The $N \times N$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\boldsymbol{\beta}_j}}$, where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\boldsymbol{\beta}_j}}[i_1, i_2] = \text{Cov}(\beta_j(\boldsymbol{s}_{i_1}), \beta_j(\boldsymbol{s}_{i_2}))$, is parameterised by an unknown vector $\boldsymbol{\theta}_{\boldsymbol{\beta}_j}$. Based on empirical variograms of the residuals from regressing the estimated $\boldsymbol{\beta}_j$ on the spatial variables $\boldsymbol{X}_j$, an exponential covariance function was used for modelling the spatial covariances $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\boldsymbol{\beta}_j}}$.

We assume the temporal patterns are captured in the mean structure through the basis functions, so the residuals $e(\boldsymbol{s}_i, t)$ are assumed to be independent over time and independent of the $\beta_j(\boldsymbol{s}_i)$. Letting $\boldsymbol{e}_t = (e(\boldsymbol{s}_1, t), \ldots, e(\boldsymbol{s}_N, t))^T$, we assume

$$\boldsymbol{e}_t \overset{\text{ind}}{\sim} N\left(0, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_e}\right), \quad t = 1, \ldots, T, \tag{5}$$

where the $N \times N$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_e}$, with $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_e}[i_1, i_2] = \text{Cov}(e(\boldsymbol{s}_{i_1}, t), e(\boldsymbol{s}_{i_2}, t))$, is parameterised by an unknown vector $\boldsymbol{\theta}_{\boldsymbol{e}}$. Similar to $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\boldsymbol{\beta}_j}}$, based on empirical variograms of the residuals, an exponential covariance function was used for modelling $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_e}$.

Given the model and the assumptions, we need to derive various covariance expressions to use in Sect. 3 when we construct improved predictions. It is not straightforward to do this but the process can be simplified if we treat the basis functions $\left\{f_j(t)\right\}_{j=0}^{J}$, once determined, as fixed because we can then formulate our hierarchical spatio-temporal model as a linear mixed model. This step (which is not included explicitly in this paper) could be used to fit the model (given the basis functions $\left\{f_j(t)\right\}_{j=0}^{J}$) in a single step using maximum likelihood or restricted maximum likelihood (REML) to improve the efficiency of the parameter estimates, although there are advantages in developing the model to use a sequential stepwise approach as in Nowak et al. (2018). We do use the linear mixed model formulation to compute the covariances needed for improved prediction; it is still complicated but it is much easier than trying to compute the covariances without this intermediate step.

The covariance between two data values $Y(\boldsymbol{s}_{i_1}, t_1)$ and $Y(\boldsymbol{s}_{i_2}, t_2)$ is given by

$$\mathrm{Cov}(Y(\boldsymbol{s}_{i_1}, t_1), Y(\boldsymbol{s}_{i_2}, t_2)) = \sum_{j=0}^{J} f_j(t_1) f_j(t_2) \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\beta_j}}[i_1, i_2] + I(t_1 = t_2)\boldsymbol{\Sigma}_{\boldsymbol{\theta}_e}[i_1, i_2].$$

Letting the N-vector $\boldsymbol{y}_t = (Y(\boldsymbol{s}_1, t), \ldots, Y(\boldsymbol{s}_N, t))^T$ denote the transpose of the tth row of the data matrix $\boldsymbol{Y}$, the covariance between $\boldsymbol{y}_{t_1}$ and $\boldsymbol{y}_{t_2}$ is given by the $N \times N$ matrix

$$\mathrm{Cov}(\boldsymbol{y}_{t_1}, \boldsymbol{y}_{t_2}) = \sum_{j=0}^{J} f_j(t_1) f_j(t_2) \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\beta_j}} + I(t_1 = t_2)\boldsymbol{\Sigma}_{\boldsymbol{\theta}_e}, \tag{6}$$

where $\mathrm{Cov}(\boldsymbol{y}_{t_1}, \boldsymbol{y}_{t_2})[i_1, i_2] = \mathrm{Cov}(Y(\boldsymbol{s}_{i_1}, t_1), Y(\boldsymbol{s}_{i_2}, t_2))$. More generally, if $\boldsymbol{y}_{t_i}^{N_i}$ denotes a generic N_i-vector of data values over N_i spatial locations at time t_i, then $\mathrm{Cov}(\boldsymbol{y}_{t_1}^{N_1}, \boldsymbol{y}_{t_2}^{N_2})$ will be the $N_1 \times N_2$ matrix given by

$$\mathrm{Cov}(\boldsymbol{y}_{t_1}^{N_1}, \boldsymbol{y}_{t_2}^{N_2}) = \sum_{j=0}^{J} f_j(t_1) f_j(t_2) \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\beta_j}\, N_1 \times N_2} + I(t_1 = t_2)\boldsymbol{\Sigma}_{\boldsymbol{\theta}_e\, N_1 \times N_2}, \tag{7}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\beta_j}\, N_1 \times N_2}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_e\, N_1 \times N_2}$ are the corresponding appropriate $N_1 \times N_2$ submatrices of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\beta_j}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_e}$, respectively. Letting the T-vector $\boldsymbol{y}_{\boldsymbol{s}_i} = (Y(\boldsymbol{s}_i, 1), \ldots, Y(\boldsymbol{s}_i, T))^T$ denote the ith column of $\boldsymbol{Y}$, the covariance between $\boldsymbol{y}_{\boldsymbol{s}_{i_1}}$ and $\boldsymbol{y}_{\boldsymbol{s}_{i_2}}$ is given by the $T \times T$ matrix

$$\mathrm{Cov}(\boldsymbol{y}_{\boldsymbol{s}_{i_1}}, \boldsymbol{y}_{\boldsymbol{s}_{i_2}}) = \sum_{j=0}^{J} \boldsymbol{f}_j \boldsymbol{f}_j^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\beta_j}}[i_1, i_2] + \boldsymbol{I}_T \boldsymbol{\Sigma}_{\boldsymbol{\theta}_e}[i_1, i_2], \tag{8}$$

where $\boldsymbol{f}_j = (f_j(1), \ldots, f_j(T))^T$, $\boldsymbol{I}_T$ is the $T \times T$ identity matrix and $\mathrm{Cov}(\boldsymbol{y}_{\boldsymbol{s}_{i_1}}, \boldsymbol{y}_{\boldsymbol{s}_{i_2}})[t_1, t_2] = \mathrm{Cov}(Y(\boldsymbol{s}_{i_1}, t_1), Y(\boldsymbol{s}_{i_2}, t_2))$. More generally, if $\boldsymbol{y}_{\boldsymbol{s}_i}^{T_i}$ denotes a generic T_i-vector of data values across T_i times points at spatial location $\boldsymbol{s}_i$, then $\mathrm{Cov}(\boldsymbol{y}_{\boldsymbol{s}_{i_1}}^{T_1}, \boldsymbol{y}_{\boldsymbol{s}_{i_2}}^{T_2})$ will be the $T_1 \times T_2$ matrix given by

$$\mathrm{Cov}(\boldsymbol{y}_{\boldsymbol{s}_{i_1}}^{T_1}, \boldsymbol{y}_{\boldsymbol{s}_{i_2}}^{T_2}) = \sum_{j=0}^{J} \boldsymbol{f}_j^{T_1} \left(\boldsymbol{f}_j^{T_2}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\beta_j}}[i_1, i_2] + \boldsymbol{I}_{T_1 \times T_2} \boldsymbol{\Sigma}_{\boldsymbol{\theta}_e}[i_1, i_2], \tag{9}$$

where $\boldsymbol{f}_j^{T_i}$ is the appropriate T_i-subvector of $\boldsymbol{f}_j$ and $\boldsymbol{I}_{T_1 \times T_2}$ is the appropriate $T_1 \times T_2$ submatrix of $\boldsymbol{I}_T$.

2.1 Variable selection on the spatial variables

Before exploring the framework for improved prediction in the hierarchical model described above, we investigated whether incorporating variable selection on the spatial variables in (4) could improve the overall model. The spatial variables available for each station were the Cartesian x and y coordinates, elevation and the climatic region to which the station belonged. The hierarchical model of Nowak et al. (2018) used the same set of spatial variables to model each $\boldsymbol{\beta}_j$. We now consider allowing the set of spatial variables to potentially differ when modelling each $\boldsymbol{\beta}_j$, using a best subsets approach to select the best model.

For each station, let X_1 and X_2 denote the x and y coordinates, respectively, X_3 denote the elevation, and I_T and I_M denote indicator variables for the top and middle regions, respectively. For each $\boldsymbol{\beta}_j$, we fitted 34 candidate models and selected the model that produced the smallest value of the Bayesian Information Criterion (BIC). These 34 candidate models were chosen based on some restrictions, namely, that the two climatic region indicator variables (I_T and I_M) must always appear together in a model and only second order interactions of each continuous variable (X_1, X_2 and X_3) with the indicator variables were considered.

For each $\boldsymbol{\beta}_j$, the selected models with the smallest BIC were the following:

$$\beta_0=\alpha_{00}+\alpha_{01}X_1+\alpha_{02}X_2+\alpha_{03}X_3+\alpha_{04}I_T+\alpha_{05}I_M+\alpha_{06}X_1I_T+\alpha_{07}X_1I_M+\epsilon \tag{10}$$

$$\beta_1=\alpha_{10}+\alpha_{11}X_1+\alpha_{12}X_2+\alpha_{13}X_3+\alpha_{14}I_T+\alpha_{15}I_M+\alpha_{16}X_1I_T+\alpha_{17}X_1I_M+\epsilon \tag{11}$$

$$\begin{aligned}\beta_2=&\alpha_{20}+\alpha_{21}X_1+\alpha_{22}X_2+\alpha_{24}I_T+\alpha_{25}I_M+\alpha_{26}X_1I_T+\alpha_{27}X_1I_M+\alpha_{28}X_2I_T\\&+\alpha_{29}X_2I_M+\epsilon\end{aligned} \tag{12}$$

$$\begin{aligned}\beta_3=&\alpha_{30}+\alpha_{31}X_1+\alpha_{32}X_2+\alpha_{33}X_3+\alpha_{34}I_T+\alpha_{35}I_M+\alpha_{36}X_1I_T+\alpha_{37}X_1I_M\\&+\alpha_{38}X_2I_T+\alpha_{39}X_2I_M+\epsilon\end{aligned} \tag{13}$$

The parameter estimates for each model, i.e., the $\hat{\boldsymbol{\alpha}}_j$, for $j=0,\ldots,J$, are displayed in Table 1. Also displayed in this table are the standard errors of each parameter estimate, which were calculated using the block bootstrap approach described in Nowak et al. (2018). We note that the standard errors displayed in Table 1 are conditional on the particular selected model. That is, for each $\boldsymbol{\beta}_j$, the same corresponding model (either (10), (11), (12) or (13)) was fitted for each bootstrap sample.

We next investigated whether this variable selection on the spatial variables leads to better predictive performance when using the estimate of the expected value $E(Y^*)$ as the predictions. We note that predictions in time use only the estimates $\hat{\boldsymbol{\beta}}_j$, whereas predictions in space use only the estimates $\hat{\boldsymbol{\alpha}}_j$. Since variable selection on the spatial variables does not affect the estimates $\hat{\boldsymbol{\beta}}_j$, predictions in time will remain unchanged with or without variable selection. As such, we will only compare predictive performance for predictions in space. Using the estimated models (10) to (13), we calculated predicted values for each station via leave-one-station-out cross-validation. The root-mean-square error (RMSE) for the predictions was 1.0804, compared to 1.0828

Table 1 Estimates of $\boldsymbol{\alpha}_j$, for $j = 0, \ldots, J$. Standard errors are given in brackets

	$j = 0$	$j = 1$	$j = 2$	$j = 3$
$\hat{\alpha}_{j0}$ (Intercept)	2.98×10^{0}	-8.15×10^{-2}	2.37×10^{-1}	2.28×10^{-1}
	(2.76×10^{-1})	(1.91×10^{-1})	(3.06×10^{-1})	(2.98×10^{-1})
$\hat{\alpha}_{j1}$ (X_1)	3.68×10^{-4}	1.58×10^{-4}	-6.88×10^{-5}	3.98×10^{-4}
	(5.13×10^{-4})	(3.20×10^{-4})	(8.70×10^{-4})	(1.48×10^{-4})
$\hat{\alpha}_{j2}$ (X_2)	-4.83×10^{-4}	5.03×10^{-4}	1.09×10^{-3}	-3.16×10^{-4}
	(4.09×10^{-4})	(2.55×10^{-4})	(7.58×10^{-4})	(1.25×10^{-4})
$\hat{\alpha}_{j3}$ (X_3)	1.06×10^{-3}	-1.90×10^{-4}		-1.44×10^{-4}
	(3.96×10^{-4})	(2.47×10^{-4})		(6.46×10^{-5})
$\hat{\alpha}_{j4}$ (I_T)	-3.09×10^{-1}	8.41×10^{-2}	-1.84×10^{-1}	2.19×10^{-1}
	(2.68×10^{-1})	(1.67×10^{-1})	(3.88×10^{-1})	(7.34×10^{-2})
$\hat{\alpha}_{j5}$ (I_M)	-2.91×10^{-1}	3.75×10^{-2}	-1.82×10^{-1}	1.96×10^{-1}
	(1.67×10^{-1})	(8.81×10^{-2})	(4.14×10^{-1})	(7.46×10^{-2})
$\hat{\alpha}_{j6}$ $(X_1 I_T)$	8.23×10^{-4}	-1.16×10^{-4}	4.32×10^{-4}	-5.50×10^{-4}
	(5.85×10^{-4})	(3.54×10^{-4})	(1.00×10^{-3})	(1.82×10^{-4})
$\hat{\alpha}_{j7}$ $(X_1 I_M)$	1.77×10^{-4}	9.10×10^{-5}	2.87×10^{-4}	-2.08×10^{-4}
	(6.52×10^{-4})	(3.86×10^{-4})	(8.70×10^{-4})	(1.55×10^{-4})
$\hat{\alpha}_{j8}$ $(X_2 I_T)$			-4.08×10^{-4}	2.00×10^{-4}
			(7.97×10^{-4})	(1.39×10^{-4})
$\hat{\alpha}_{j9}$ $(X_2 I_M)$			-4.73×10^{-4}	4.22×10^{-4}
			(1.11×10^{-3})	(1.87×10^{-4})

without variable selection. We see that variable selection provides a very marginal improvement in predictive performance when predicting in space. As a visual comparison, the predicted and observed values, averaged over time, are displayed in Fig. 1 (no variable selection on the left, variable selection on the right). We see that the more accurate predictions for stations in the south-east region (blue circles with smaller radii) may be driving the lower overall RMSE. The selected models for $\boldsymbol{\beta}_j$ given in (10) to (13) will be used in Sects. 3.2 and 3.3 for predicting in space and predicting in both space and time, respectively.

3 Applying the framework for improved prediction

Suppose we wish to predict a new data value $Y(\boldsymbol{s}^*, t^*)$. From (2), the improved prediction $\hat{Y}(\boldsymbol{s}^*, t^*)$ is given by

$$\hat{Y}(\boldsymbol{s}^*, t^*) = \hat{E}(Y(\boldsymbol{s}^*, t^*)) + \widehat{\mathrm{Cov}}(Y(\boldsymbol{s}^*, t^*), \boldsymbol{y}^{\mathrm{tr}})\widehat{\mathrm{Var}}(\boldsymbol{y}^{\mathrm{tr}})^{-1}\left(\boldsymbol{y}^{\mathrm{tr}} - \hat{E}(\boldsymbol{y}^{\mathrm{tr}})\right). \quad (14)$$

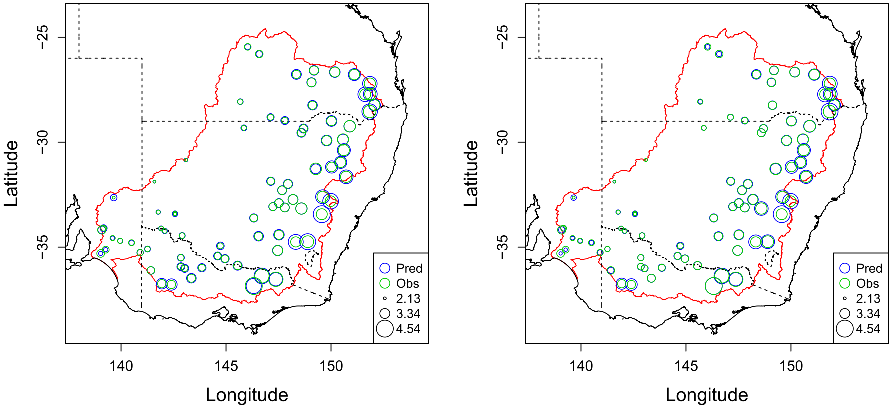


Fig. 1 Predicted (blue) and observed (green) values, averaged over time, for each station. The left plot displays the predicted values without variable selection and the right plot displays the predicted values with variable selection on the spatial variables

To apply these improved predictions to the hierarchical model described in Sect. 2, we will need to calculate the appropriate expected value, covariance and variance terms in (14). How these terms are calculated from the hierarchical model will depend on the type of prediction desired. We will focus on three types of prediction: predicting at an unobserved time point, predicting at an unobserved spatial location, and predicting at an unobserved time point at an unobserved spatial location. For each type of prediction, we set aside some of the data to serve as a test data set on which to compare the improved predictions of (14) to predictions using only $\hat{E}(Y(\boldsymbol{s}^*, t^*))$.

We noted in Sect. 1 that calculating the covariance and variance terms in (14) can be very computationally intensive for large data sets. As such, we explored the use of reduced subsets of the data to compute these terms. Further, we investigated how the prediction accuracy was affected by varying the size of the subsets. For each type of prediction, we considered three different methods of subsetting the data, which are described schematically in Figs. 2, 4 and 6. Note that the reduced subsets were only used for computing the covariance and variance terms. Any model parameters that are required to calculate the expected value, covariance and variance terms in (14), that is, $\left\{\boldsymbol{\beta}_j\right\}_{j=0}^{J}$, $\left\{\boldsymbol{\alpha}_j\right\}_{j=0}^{J}$, $\left\{\boldsymbol{\theta}_{\boldsymbol{\beta}_j}\right\}_{j=0}^{J}$ and $\boldsymbol{\theta}_{\boldsymbol{e}}$, were estimated using the full training data.

3.1 Prediction in time

Prediction in time refers to the problem of predicting the data value at a future time point for an existing or observed spatial location. That is, we are wanting to predict $Y^{\text{tst}} = Y(\boldsymbol{s}_i, k)$, for $i = 1, \ldots, N$, for some future time point k. Therefore, we will set aside the most recent $T^{\text{tst}} = 12$ months of the data as the test data and use the remaining $T^{\text{tr}} = 974$ months as the training data. That is, $\boldsymbol{Y}^{\text{tst}}$ will be the $T^{\text{tst}} \times N$ matrix consisting of the last 12 rows of $\boldsymbol{Y}$ and $\boldsymbol{Y}^{\text{tr}}$ will be the $T^{\text{tr}} \times N$ matrix consisting of the first 974 rows of $\boldsymbol{Y}$.

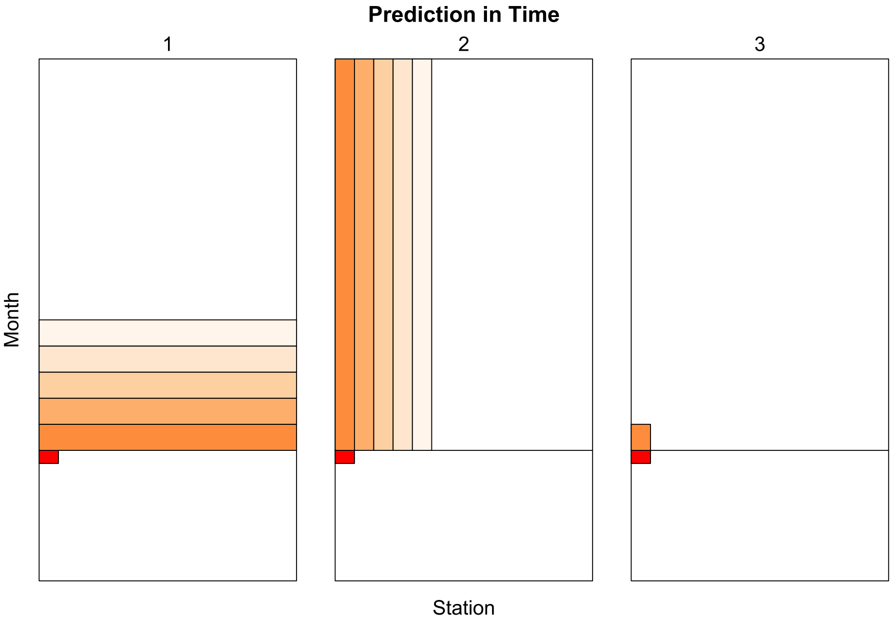


Fig. 2 Three methods of subsetting the training data for computing the covariance and variance terms when predicting in time. Each rectangle represents the data matrix $\boldsymbol{Y}$ with months in the rows and stations in the columns. The test data to predict (at a future month) is indicated in red. Moving from left to right across the columns, the stations are assumed to be ordered in increasing distance from the test station. The first method subsets the most recent months across all stations. The second method subsets the test and nearby stations for all months. The third method subsets the most recent months for the test station

The improved prediction $\hat{Y}(\boldsymbol{s}_i, k)$ is calculated using (14), noting that

$$\hat{E}(Y(\boldsymbol{s}_i, k)) = \sum_{j=0}^{J} \hat{\beta}_j(\boldsymbol{s}_i) f_j(k), \tag{15}$$

where we use the approach described in Nowak et al. (2018) to calculate the value of the basis functions $f_j(t)$ at the future time point k. Specifically, the deterministic basis functions $f_1(t)$ and $f_2(t)$ were extended in the natural way. The empirically derived basis function $f_3(t)$ was extended by setting the value at each future month to be the value at that month in the previous year. Predictions using (15) resulted in an RMSE of 0.9190 on the test data, which we will use as a baseline for comparison with our improved predictions. To feasibly compute $\widehat{\mathrm{Cov}}(Y(\boldsymbol{s}_i, k), \boldsymbol{y}^{\mathrm{tr}})$ and $\widehat{\mathrm{Var}}(\boldsymbol{y}^{\mathrm{tr}})$ in (14), we will investigate using reduced subsets of the training data. Figure 2 presents a graphical representation of the three subsetting methods used.

For subsetting in time, we selected the most recent l months across all N stations, i.e., the last l rows of $\boldsymbol{Y}^{\mathrm{tr}}$. Specifically, we will set the vectorised subsetted training data to be the Nl-vector $\boldsymbol{y}^{\mathrm{tr}} = ((\boldsymbol{y}^{\mathrm{tr}}_{T^{\mathrm{tr}}-l+1})^T, \ldots, (\boldsymbol{y}^{\mathrm{tr}}_{T^{\mathrm{tr}}})^T)^T$, where $\boldsymbol{y}^{\mathrm{tr}}_t$ denotes the transpose of the tth row of $\boldsymbol{Y}^{\mathrm{tr}}$. Since the same training data is used for predicting each station in the test data, we can efficiently calculate the predictions across all stations in one step. Letting $\boldsymbol{y}_k^{\text{tst}}$ denote the kth row of $\boldsymbol{Y}^{\text{tst}}$, the improved predictions for $\boldsymbol{y}_k^{\text{tst}}$ are given by

$$\hat{\boldsymbol{y}}_k^{\text{tst}} = \hat{E}(\boldsymbol{y}_k^{\text{tst}}) + \widehat{\text{Cov}}(\boldsymbol{y}_k^{\text{tst}}, \boldsymbol{y}^{\text{tr}})\widehat{\text{Var}}(\boldsymbol{y}^{\text{tr}})^{-1}\left(\boldsymbol{y}^{\text{tr}} - \hat{E}(\boldsymbol{y}^{\text{tr}})\right), \tag{16}$$

where

$$\begin{aligned}
\hat{E}(\boldsymbol{y}_k^{\text{tst}}) &= \sum_{j=0}^{J} \hat{\boldsymbol{\beta}}_j f_j(k), \\
\widehat{\text{Cov}}(\boldsymbol{y}_k^{\text{tst}}, \boldsymbol{y}^{\text{tr}}) &= \left[\widehat{\text{Cov}}(\boldsymbol{y}_k^{\text{tst}}, \boldsymbol{y}^{\text{tr}}_{T^{\text{tr}}-l+1}) \dots \widehat{\text{Cov}}(\boldsymbol{y}_k^{\text{tst}}, \boldsymbol{y}^{\text{tr}}_{T^{\text{tr}}})\right] \quad \text{and} \\
\widehat{\text{Var}}(\boldsymbol{y}^{\text{tr}}) &= \begin{bmatrix} \widehat{\text{Cov}}(\boldsymbol{y}^{\text{tr}}_{T^{\text{tr}}-l+1}, \boldsymbol{y}^{\text{tr}}_{T^{\text{tr}}-l+1}) & \dots & \widehat{\text{Cov}}(\boldsymbol{y}^{\text{tr}}_{T^{\text{tr}}-l+1}, \boldsymbol{y}^{\text{tr}}_{T^{\text{tr}}}) \\ \vdots & \ddots & \vdots \\ \widehat{\text{Cov}}(\boldsymbol{y}^{\text{tr}}_{T^{\text{tr}}}, \boldsymbol{y}^{\text{tr}}_{T^{\text{tr}}-l+1}) & \dots & \widehat{\text{Cov}}(\boldsymbol{y}^{\text{tr}}_{T^{\text{tr}}}, \boldsymbol{y}^{\text{tr}}_{T^{\text{tr}}}) \end{bmatrix}.
\end{aligned}$$

The individual covariance terms are calculated using (6). Using (16), predictions for the test data were produced for $l \in \{12, 24, 36, 48, 60\}$. The RMSE for each value of l is displayed (green line) in Fig. 3. We see that there is a small improvement in RMSE compared to predictions using (15). Further, subsetting the most recent four years (48 months) of the training data seems to be optimal, resulting in an RMSE of 0.9113.

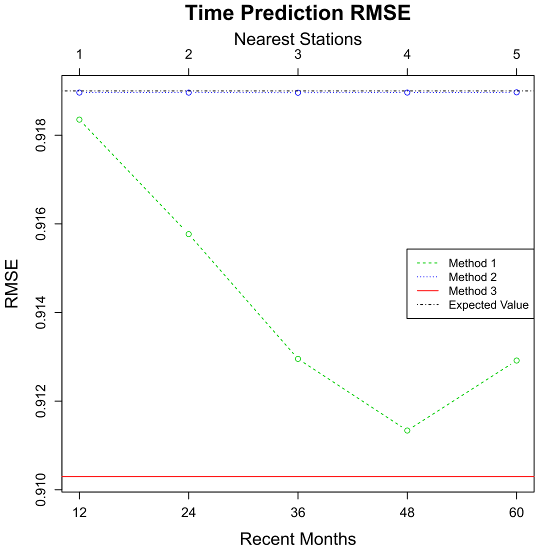
For subsetting in space, for a given station, we selected all months from a set of neighbouring stations. That is, we subset $\boldsymbol{Y}^{\text{tr}}$ by selecting the columns corresponding to the set of neighbouring stations. Since the subsets vary for each station, predictions need to be produced separately for each station. For the ith station $\boldsymbol{s}_i$, let $S_i^m = \{i_1, \dots, i_m\}$ denote the set of indices for the m nearest stations by distance (including the ith station) and set the vectorised subsetted training data to be the mT^{tr}-vector $\boldsymbol{y}^{\text{tr}} = ((\boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_{i_1}})^T, \dots, (\boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_{i_m}})^T)^T$, where $\boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_n}$ denotes the nth column of $\boldsymbol{Y}^{\text{tr}}$. Predictions for all months in the test data can then be efficiently calculated in one step. Letting $\boldsymbol{y}^{\text{tst}}_{\boldsymbol{s}_i}$ denote the ith column of $\boldsymbol{Y}^{\text{tst}}$, the improved predictions are given by

$$\hat{\boldsymbol{y}}_{\boldsymbol{s}_i}^{\text{tst}} = \hat{E}(\boldsymbol{y}_{\boldsymbol{s}_i}^{\text{tst}}) + \widehat{\text{Cov}}(\boldsymbol{y}_{\boldsymbol{s}_i}^{\text{tst}}, \boldsymbol{y}^{\text{tr}})\widehat{\text{Var}}(\boldsymbol{y}^{\text{tr}})^{-1}\left(\boldsymbol{y}^{\text{tr}} - \hat{E}(\boldsymbol{y}^{\text{tr}})\right), \tag{17}$$

where

$$\begin{aligned}
\hat{E}(\boldsymbol{y}_{\boldsymbol{s}_i}^{\text{tst}}) &= \sum_{j=0}^{J} \hat{\beta}_j(\boldsymbol{s}_i) \boldsymbol{f}_j^{T^{\text{tst}}}, \\
\widehat{\text{Cov}}(\boldsymbol{y}_{\boldsymbol{s}_i}^{\text{tst}}, \boldsymbol{y}^{\text{tr}}) &= \left[\widehat{\text{Cov}}(\boldsymbol{y}_{\boldsymbol{s}_i}^{\text{tst}}, \boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_{i_1}}) \dots \widehat{\text{Cov}}(\boldsymbol{y}_{\boldsymbol{s}_i}^{\text{tst}}, \boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_{i_m}})\right] \quad \text{and} \\
\widehat{\text{Var}}(\boldsymbol{y}^{\text{tr}}) &= \begin{bmatrix} \widehat{\text{Cov}}(\boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_{i_1}}, \boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_{i_1}}) & \dots & \widehat{\text{Cov}}(\boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_{i_1}}, \boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_{i_m}}) \\ \vdots & \ddots & \vdots \\ \widehat{\text{Cov}}(\boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_{i_m}}, \boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_{i_1}}) & \dots & \widehat{\text{Cov}}(\boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_{i_m}}, \boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_{i_m}}) \end{bmatrix}.
\end{aligned}$$

Fig. 3 RMSE over the test data for prediction in time for different methods of subsetting the training data. The green, blue and red lines denote subsetting methods 1, 2, and 3, respectively, of Fig. 2. The black line (which is just above the blue line) denotes predictions using (15)



The individual covariance terms are calculated using (8) and (9). Using (17), predictions for the test data were produced for $m \in \{1, 2, 3, 4, 5\}$. The RMSE for each value of m is displayed (blue line) in Fig. 3. The RMSE was mostly unchanged across different values of m, with the minimum of 0.9190 being achieved when $m = 1$. The RMSEs were also quite similar to using (15) for predictions.

We can draw some insights from the results of these two methods of subsetting the training data. In terms of subsetting in time, the greatest gains in prediction accuracy occur when selecting a recent set of months (e.g., 4 years) rather than all months in the training data. In terms of subsetting in space, there does not appear to be much gain from including any neighbouring stations in addition to the station itself. Combining these observations, the optimal subset of the training data to use when predicting a given station's test data may simply be a selection of recent months from only the station itself.

Further to the above, for a given station, we subsetted the training data in both time and space by selecting the most recent $l = 48$ months from the given station. Specifically, for the ith station $\boldsymbol{s}_i$, we set the vectorised subsetted training data to be the l-vector $\boldsymbol{y}^{\text{tr}} = (Y^{\text{tr}}(\boldsymbol{s}_i, T^{\text{tr}} - l + 1), \ldots, Y^{\text{tr}}(\boldsymbol{s}_i, T^{\text{tr}}))^T$. We can then use (17) for prediction, with $\widehat{\text{Cov}}(\boldsymbol{y}^{\text{tst}}_{\boldsymbol{s}_i}, \boldsymbol{y}^{\text{tr}})$ and $\widehat{\text{Var}}(\boldsymbol{y}^{\text{tr}})$ now calculated using (9). Using this method for subsetting the training data resulted in an RMSE of 0.9103 (red line in Fig. 3), which was the lowest among all subsetting methods.

3.2 Prediction in space

Prediction in space refers to the problem of predicting the data value at an unobserved spatial location at observed time points. The problem can effectively be thought of as spatial interpolation. Specifically, we want to predict $Y^{\text{tst}} = Y(\boldsymbol{s}_i, t)$, for $t =$

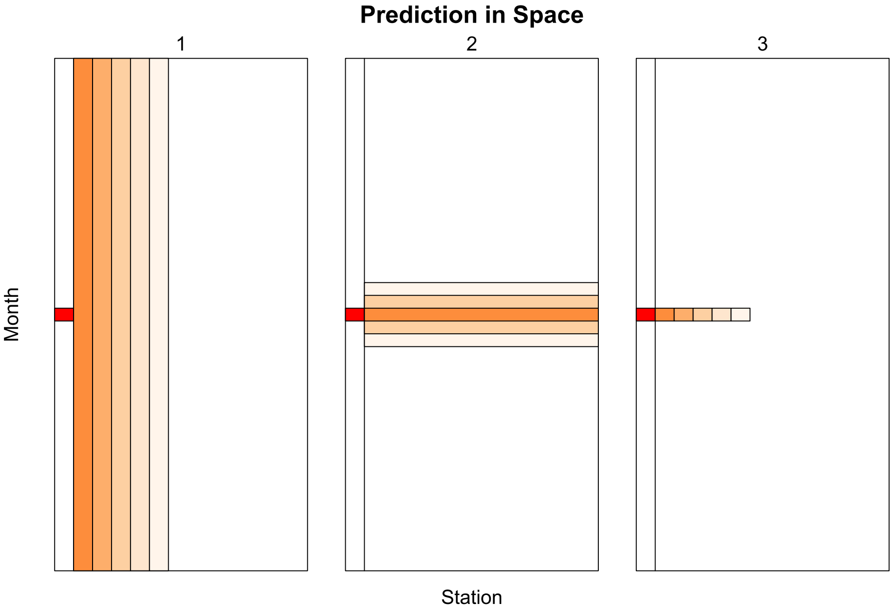


Fig. 4 Three methods of subsetting the training data for computing the covariance and variance terms when predicting in space. Similar to Fig. 2, each rectangle represents the data matrix $\boldsymbol{Y}$ and the test data to predict (at an unobserved station) is indicated in red. The first method subsets the nearby stations for all months. The second method subsets a window of months around the test month for all stations. The third method subsets nearby stations for the test month

$1, \ldots, T$, at some unobserved spatial location $\boldsymbol{s}_i$. We will use leave-one-station-out cross-validation to evaluate the performance of our improved prediction in space. That is, each station will serve as a "test station". Therefore, for each $i = 1, \ldots, N$, the test data $\boldsymbol{Y}^{\text{tst}} = \boldsymbol{y}^{\text{tst}}_{\boldsymbol{s}_i}$ will be the T-vector that is the ith column of $\boldsymbol{Y}$ and the training data $\boldsymbol{Y}^{\text{tr}}$ will be the $T \times (N-1)$ matrix consisting of the remaining columns of $\boldsymbol{Y}$.

The improved prediction $\hat{Y}(\boldsymbol{s}_i, t)$ is calculated according to (14), with

$$\hat{E}(Y(\boldsymbol{s}_i, t)) = \sum_{j=0}^{J} \hat{\beta}_j(\boldsymbol{s}_i) f_j(t) = \sum_{j=0}^{J} \boldsymbol{x}_{j,i}^T \hat{\boldsymbol{\alpha}}_j f_j(t), \tag{18}$$

where $\boldsymbol{x}_{j,i}$ denotes the vector of spatial variables for location $\boldsymbol{s}_i$. Note that since $\boldsymbol{s}_i$ is now an unobserved location, the estimates $\{\hat{\beta}_j(\boldsymbol{s}_i)\}_{j=0}^{J}$ are unknown and need to be calculated according to (10) to (13). Predictions using (18) resulted in an RMSE of 1.0804 on the test data. As a naive comparison, when using the nearest station's observed data values as the predictions for each test station, the RMSE was 0.6478. The naive predictions outperforming predictions using (18) indicates there is potential for the improved predictions of (14) to substantially increase predictive accuracy. For the improved predictions, we will again investigate using reduced subsets of the training data, with the details described in Fig. 4.

To subset in space, for each test station, we selected all months from a set of neighbouring stations. That is, for the ith test station $\boldsymbol{s}_i$, let $S_i^m = \{i_1, \dots, i_m\}$ denote the set of indices for the m nearest stations by distance in the training data. The vectorised subsetted training data is then set to be the mT-vector $\boldsymbol{y}^{\mathrm{tr}} = ((\boldsymbol{y}^{\mathrm{tr}}_{\boldsymbol{s}_{i_1}})^T, \dots, (\boldsymbol{y}^{\mathrm{tr}}_{\boldsymbol{s}_{i_m}})^T)^T$, where $\boldsymbol{y}^{\mathrm{tr}}_{\boldsymbol{s}_n}$ denotes the nth column of $\boldsymbol{Y}^{\mathrm{tr}}$. Improved predictions over all months in the test station, i.e., $\boldsymbol{y}^{\mathrm{tst}}_{\boldsymbol{s}_i}$, are calculated according to (17), but now with $\hat{E}(\boldsymbol{y}^{\mathrm{tst}}_{\boldsymbol{s}_i}) = \sum_{j=0}^{J} \boldsymbol{x}^T_{j,i} \hat{\boldsymbol{\alpha}}_j \boldsymbol{f}_j$. Predictions for the test data were produced for $m \in \{1, 2, 3, 4, 5\}$. The RMSE for each value of m is displayed (dark blue line) in Fig. 5. Even for $m = 1$, we see that the RMSE is lower than both the predictions using only (18) and the naive predictions. As we include more neighbouring stations, the RMSE continues to decrease, reaching a minimum of 0.5588.

To subset in time, for each month in the test station, we selected all stations for a window of months surrounding this month of interest. Since we are predicting at observed time points, we are able to consider data from both previous, current and future months, relative to the month of interest, for subsetting the training data. Specifically, for month t, let $T_t^l = \{z \in \mathbb{Z} : \max\{t-l, 1\} \leq z \leq \min\{t+l, T\}\}$, where $l \ll T$, denote the set of indices corresponding to the window of $|T_t^l| = \min\{t+l, T\} - \max\{t-l, 1\} + 1$ months centered at month t. Note that the window size ranges from a maximum of $2l+1$ months (when $l+1 \leq t \leq T-l$) to a minimum of $l+1$ months (when $t = 1$ and T). The vectorised subsetted training data is therefore set to be the $(N-1)|T_t^l|$-vector $\boldsymbol{y}^{\mathrm{tr}} = ((\boldsymbol{y}^{\mathrm{tr}}_{\max\{t-l,1\}})^T, \dots, (\boldsymbol{y}^{\mathrm{tr}}_{\min\{t+l,T\}})^T)^T$. Since the vectorised subsetted training data will depend on the month t, for the ith test station $\boldsymbol{s}_i$ we calculated predictions separately for each month t using (14), where now $\widehat{\mathrm{Cov}}(Y(\boldsymbol{s}_i, t), \boldsymbol{y}^{\mathrm{tr}}) = \left[\widehat{\mathrm{Cov}}(Y(\boldsymbol{s}_i, t), \boldsymbol{y}^{\mathrm{tr}}_{\max\{t-l,1\}}) \dots \widehat{\mathrm{Cov}}(Y(\boldsymbol{s}_i, t), \boldsymbol{y}^{\mathrm{tr}}_{\min\{t+l,T\}})\right]$ and

$$\widehat{\mathrm{Var}}(\boldsymbol{y}^{\mathrm{tr}}) = \begin{bmatrix} \widehat{\mathrm{Cov}}(\boldsymbol{y}^{\mathrm{tr}}_{\max\{t-l,1\}}, \boldsymbol{y}^{\mathrm{tr}}_{\max\{t-l,1\}}) & \dots & \widehat{\mathrm{Cov}}(\boldsymbol{y}^{\mathrm{tr}}_{\max\{t-l,1\}}, \boldsymbol{y}^{\mathrm{tr}}_{\min\{t+l,T\}}) \\ \vdots & \ddots & \vdots \\ \widehat{\mathrm{Cov}}(\boldsymbol{y}^{\mathrm{tr}}_{\min\{t+l,T\}}, \boldsymbol{y}^{\mathrm{tr}}_{\max\{t-l,1\}}) & \dots & \widehat{\mathrm{Cov}}(\boldsymbol{y}^{\mathrm{tr}}_{\min\{t+l,T\}}, \boldsymbol{y}^{\mathrm{tr}}_{\min\{t+l,T\}}) \end{bmatrix}.$$

The individual covariance terms were calculated using (6) and (7). Predictions for the test data were produced for $l \in \{0, 1, 2, 3, 4\}$. The RMSE for each value of l is displayed (green line) in Fig. 5. The minimum RMSE of approximately 0.5526 was obtained for $l = 0$. This was similar to the minimum RMSE achieved when subsetting in space.

These results indicate that, when subsetting the training data in space, including more neighbouring stations improves prediction accuracy and, when subsetting in time, only training data from the month of interest are required. To explore this further, we subsetted the training data in both time and space. In detail, for each month in the test station, we selected a set of neighbouring stations only for this month of interest. Similar to the setup used for subsetting in space, for the ith test station $\boldsymbol{s}_i$, let $S_i^m = \{i_1, \dots, i_m\}$ denote the set of indices for the m nearest stations by distance in the training data. For month t, the vectorised subsetted training data is set to be the m-vector $\boldsymbol{y}^{\mathrm{tr}} = (Y^{\mathrm{tr}}(\boldsymbol{s}_{i_1}, t), \dots, Y^{\mathrm{tr}}(\boldsymbol{s}_{i_m}, t))^T$. For the ith test station $\boldsymbol{s}_i$,

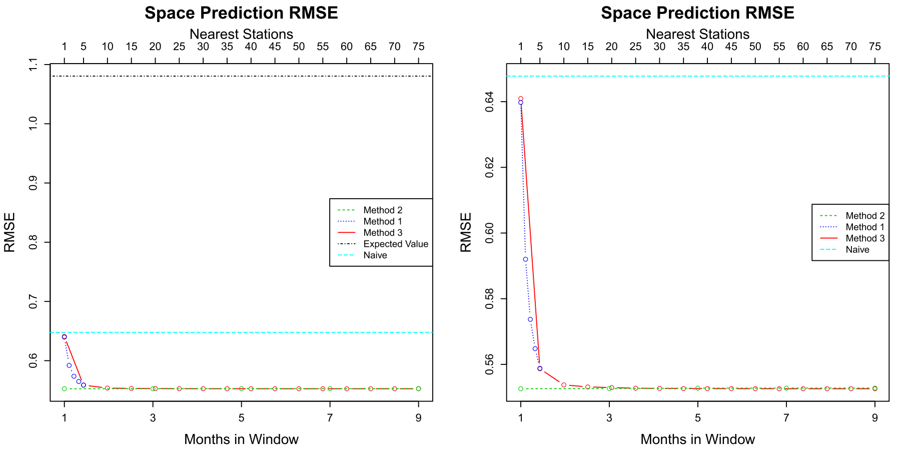


Fig. 5 RMSE via leave-one-station-out cross-validation for prediction in space for different methods of subsetting the training data. The dark blue, green and red lines denote subsetting methods 1, 2 and 3, respectively, of Fig. 4. The black line denotes predictions using (18) and the light blue line denotes the naive prediction. The right plot is a zoomed view of the lower region of the left plot

predictions were calculated separately for each month t again using (14), noting that the calculations of $\widehat{\text{Cov}}(Y(\boldsymbol{s}_i, t), \boldsymbol{y}^{\text{tr}})$ and $\widehat{\text{Var}}(\boldsymbol{y}^{\text{tr}})$ will now be somewhat simplified compared to when subsetting in time. Predictions for the test data were produced for $m \in \{1, 5, 10, 15, \ldots, 75\}$. The RMSE for each value of m is displayed (red line) in Fig. 5. While the minimum RMSE was achieved at $m = 60$ (0.5526), the RMSE was approximately 0.553 from $m = 15$ onwards. Therefore, for each month in the test station, optimal prediction accuracy can be achieved by subsetting on only 15 data values in the training data.

3.3 Prediction in space and time

For spatio-temporal data, prediction in both space and time is the most challenging problem but is often of most interest. It involves predicting the data value $Y^{\text{tst}} = Y(\boldsymbol{s}_i, k)$ for some future time point k at an unobserved spatial location $\boldsymbol{s}_i$. For each $i = 1, \ldots, N$, the test data $\boldsymbol{Y}^{\text{tst}}$ will be the vector of length $T^{\text{tst}} = 12$ corresponding to the last 12 rows (months) of the ith column (station) of $\boldsymbol{Y}$ and the training data $\boldsymbol{Y}^{\text{tr}}$ will be the $T^{\text{tr}} \times (N - 1)$ matrix consisting of the first $T^{\text{tr}} = 974$ rows of $\boldsymbol{Y}$ with the ith column removed.

The improved prediction $\hat{Y}(\boldsymbol{s}_i, k)$ is calculated using (14), where now

$$\hat{E}(Y(\boldsymbol{s}_i, k)) = \sum_{j=0}^{J} \hat{\beta}_j(\boldsymbol{s}_i) f_j(k) = \sum_{j=0}^{J} \boldsymbol{x}_{j,i}^T \hat{\boldsymbol{\alpha}}_j f_j(k). \tag{19}$$

Note that the estimates $\{\hat{\beta}_j(\boldsymbol{s}_i)\}_{j=0}^{J}$ need to be calculated according to (10) to (13), as was done when predicting in space, and the values of the basis functions $f_j(t)$

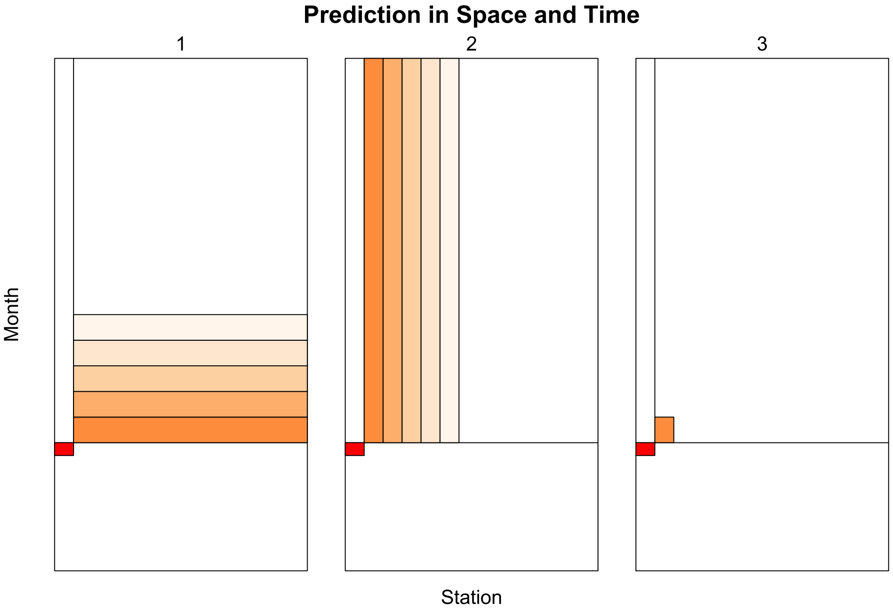


Fig. 6 Three methods of subsetting the training data for computing the covariance and variance terms when predicting in space and time. The subsetting methods are the same as that used for prediction in time (Fig. 2), with the only difference being that data in the test station are no longer included in the training data. This is because we are now predicting at a future month at an unobserved station

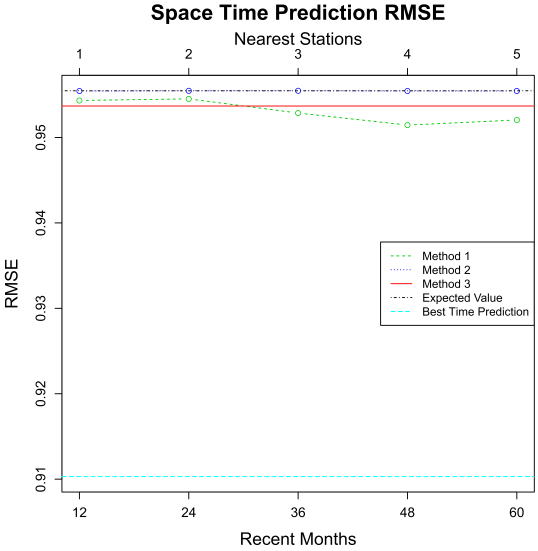
at the future time point k are calculated similarly to when predicting in time. Given that we are using the same test data used when predicting in time, we will follow a similar method for subsetting the training data. We will use the RMSE obtained when using (19) for prediction (0.9555) and the minimum RMSE achieved when predicting in time (0.9103) as baseline comparisons. The subsetting methods are described in Fig. 6.

For subsetting in time, we selected the most recent l months across the $N-1$ stations in the training data. Hence we set the vectorised subsetted training data to be the $(N-1)l$-vector $\boldsymbol{y}^{\mathrm{tr}} = ((\boldsymbol{y}^{\mathrm{tr}}_{T^{\mathrm{tr}}-l+1})^T, \dots, (\boldsymbol{y}^{\mathrm{tr}}_{T^{\mathrm{tr}}})^T)^T$. For each $i = 1, \dots, N$, the improved predictions for all months in the test station, i.e., $\boldsymbol{y}^{\mathrm{tst}}_{\boldsymbol{s}_i}$, are calculated according to (17), where $\hat{E}(\boldsymbol{y}^{\mathrm{tst}}_{\boldsymbol{s}_i}) = \sum_{j=0}^{J} \boldsymbol{x}^T_{j,i} \hat{\boldsymbol{\alpha}}_j \boldsymbol{f}^{T^{\mathrm{tst}}}_j$,

$$\widehat{\mathrm{Cov}}(\boldsymbol{y}^{\mathrm{tst}}_{\boldsymbol{s}_i}, \boldsymbol{y}^{\mathrm{tr}}) = \begin{bmatrix} \widehat{\mathrm{Cov}}(Y^{\mathrm{tst}}(\boldsymbol{s}_i, 1), \boldsymbol{y}^{\mathrm{tr}}_{T^{\mathrm{tr}}-l+1}) & \dots & \widehat{\mathrm{Cov}}(Y^{\mathrm{tst}}(\boldsymbol{s}_i, 1), \boldsymbol{y}^{\mathrm{tr}}_{T^{\mathrm{tr}}}) \\ \vdots & \ddots & \vdots \\ \widehat{\mathrm{Cov}}(Y^{\mathrm{tst}}(\boldsymbol{s}_i, T^{\mathrm{tst}}), \boldsymbol{y}^{\mathrm{tr}}_{T^{\mathrm{tr}}-l+1}) & \dots & \widehat{\mathrm{Cov}}(Y^{\mathrm{tst}}(\boldsymbol{s}_i, T^{\mathrm{tst}}), \boldsymbol{y}^{\mathrm{tr}}_{T^{\mathrm{tr}}}) \end{bmatrix}$$

and $\widehat{\mathrm{Var}}(\boldsymbol{y}^{\mathrm{tr}})$ is calculated as was done in (16). The individual covariance terms are calculated using (6) and (7). Predictions for the test data were produced for $l \in \{12, 24, 36, 48, 60\}$. The RMSE for each value of l is displayed (green line) in Fig. 7.

Fig. 7 RMSE over the test data for prediction in both space and time for different methods of subsetting the training data. The green, blue and red lines denote subsetting methods 1, 2 and 3, respectively, of Fig. 6. The black line (which is almost on top of the dark blue line) denotes predictions using (19) and the light blue line denotes the best predictions in time of Sect. 3.1



Similar to what was observed when predicting in time, the minimum RMSE of 0.9515 was achieved when $l = 48$.

For subsetting in space, for a given station, we selected all months from a set of neighbouring stations. For the ith station $\boldsymbol{s}_i$, let $S_i^m = \{i_1, \ldots, i_m\}$ denote the set of indices for the m nearest stations by distance in the training data. The vectorised subsetted training data is then the mT^{tr}-vector $\boldsymbol{y}^{\text{tr}} = ((\boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_{i_1}})^T, \ldots, (\boldsymbol{y}^{\text{tr}}_{\boldsymbol{s}_{i_m}})^T)^T$. For each $i = 1, \ldots, N$, improved predictions for all months in the test station are again computed using (17), but now with $\widehat{\text{Cov}}(\boldsymbol{y}^{\text{tst}}_{\boldsymbol{s}_i}, \boldsymbol{y}^{\text{tr}})$ and $\widehat{\text{Var}}(\boldsymbol{y}^{\text{tr}})$ calculated in a similar manner to (17). Predictions for the test data were produced for $m \in \{1, 2, 3, 4, 5\}$. The RMSE for each value of m is displayed (blue line) in Fig. 7. Again, as was observed when predicting in time, the RMSE remained mostly unchanged across different values of m, with the minimum (0.9554) being achieved when $m = 1$.

As a final comparison, we subsetted the training data in both time and space. Based on the results for subsetting in time and subsetting in space, for a given station, we selected the most recent $l = 48$ months for the nearest station in the training data (i.e., $\boldsymbol{s}_{i_1}$). The vectorised subsetted training data is then the l-vector $\boldsymbol{y}^{\text{tr}} = (Y^{\text{tr}}(\boldsymbol{s}_{i_1}, T^{\text{tr}} - l + 1), \ldots, Y^{\text{tr}}(\boldsymbol{s}_{i_1}, T^{\text{tr}}))^T$. For each $i = 1, \ldots, N$, (17) is again used to produce improved predictions for all months in the test station, where $\widehat{\text{Cov}}(\boldsymbol{y}^{\text{tst}}_{\boldsymbol{s}_i}, \boldsymbol{y}^{\text{tr}})$ and $\widehat{\text{Var}}(\boldsymbol{y}^{\text{tr}})$ are now calculated using (9). The RMSE was 0.9537 and is displayed in red in Fig. 7.

Comparing these different methods of subsetting the training data, we see that the greatest prediction accuracy on the test data is attained when we subset on all stations for the most recent 48 months in the training data. However, using only the nearest station for the most recent 48 months produces a similar RMSE, indicating that it may not be necessary to subset on all stations in the training data. We note that regardless of the method used for subsetting the training data, the RMSE on the test data when predicting in both space and time is greater than the RMSE when predicting only in time. This is expected since, for each station in the test data, when predicting only in time we have the advantage of being able to use any predictive power contained in past data observed at this station.

4 Discussion

Predicting a new data value from a parametric model typically involves using an estimate of the expectation of the new data value as the predicted value. We have applied a framework for prediction for models for spatio-temporal data that improves on using only the expectation of the value to be predicted. A key advantage of this framework is that it is a general framework that can be applied to many parametric models for spatio-temporal data. The improved predictions were based on minimising the mean squared prediction error. The improved predictions have a strong optimality property under normality and a weaker optimality property when normality does not hold. As such, the framework has the potential to improve predictions over simply using the estimated mean regardless of the distributional assumptions on the data.

We applied the framework for improving predictions to a hierarchical model for monthly rainfall data in the Murray-Darling Basin. We focused on all the main types of prediction for spatio-temporal data, namely, prediction only in time, prediction only in space and prediction in both time and space. In all situations, the framework improved prediction accuracy compared to using only the expectation of the value to be predicted. In particular, the improvement in predictive accuracy was quite large when predicting only in space. This confirms the value of kriging for spatial prediction.

One potential limitation of the framework is that the calculations of the improved predictions may be computationally intensive. This is mainly due to the need to calculate or invert large covariance matrices when generating the predictions. However, we have demonstrated that large gains in predictive accuracy can still be achieved by using only a reduced subset of the data for computing these matrices. Specifically, for our model, we obtained good results for prediction in

- Time using only the previous 48 months data from the station of interest;
- Space using only data from 15 neighbouring stations for only the month of interest;
- Space and time using data from all stations for only the most recent 48 months, although there is some evidence that fewer stations can be used.

The results for prediction in time are interesting because they are similar to what we expect when using an autoregressive time series model (for which we only need a small fixed number of lagged values for prediction) even though our model captures temporal effects through the basis functions $\left\{f_j(t)\right\}_{j=0}^{J}$ rather than through an autoregressive model. The results for prediction in space are similar to what we expect from kriging in spatial models. We believe that these results show that the framework provides a computationally feasible approach for improving predictions that should be applied to a wide variety of models for spatio-temporal data.

We have presented and evaluated the predictions made in this paper as point predictions. The assessment of uncertainty for these predictions using analytical approximations for the mean squared prediction errors or the bootstrap is quite complicated. A useful recent paper on the topic in the spatial context (Thiart et al. 2014) provides some optimism that bootstrap methods may be useful in our context. This is attractive because we have already developed a block bootstrap method to assess the uncertainty in the parameter estimates. The application of the method to estimating prediction uncertainty requires further investigation.

Acknowledgements This research was partially supported under the Australian Research Council's Discovery Projects funding scheme (project number DP180100836). We are grateful to Ichiro Ken Shimatani and two anonymous referees for their comments.

References

Banerjee S, Carlin BP, Gelfand AE (2015) Hierarchical modeling and analysis for spatial data, 2nd edn. CRC Press, New York

Cressie N (1990) The origins of kriging. Math Geol 22(3):239–252

Harville D (1976) Extension of the Gauss-Markov Theorem to include the estimation of random effects. Ann Stat 4(2):384–395

Henderson CR (1950) Estimation of genetic parameters. Ann Math Stat 21(2):309–310

Krige DG (1951) A statistical approach to some basic mine valuation problems on the Witwatersrand. J Chem Metallurg Mining Soc South Afr 52(6):119–139

Krige DG (1962) Effective pay limits for selective mining. J South Afr Inst Mining Metallurgy 62(6):345–363

Nowak G, Welsh AH, O'Neill TJ, Feng L (2018) Spatio-temporal modelling of rainfall in the Murray-Darling Basin. J Hydrol 557:522–538

Robinson GK (1991) That BLUP is a good thing: the estimation of random effects. Stat Sci 6(1):15–32

Thiart C, Ngwenya MZ, Haines LM (2014) Investigating 'optimal' kriging variance estimation using an analytic and a bootstrap approach. J South Afr Inst Mining Metallurgy 114(8):613–618

Gen Nowak is currently a senior lecturer in the Research School of Finance, Actuarial Studies and Statistics at the Australian National University. Prior to joining the ANU, he completed his PhD in Statistics at Stanford University and was a postdoctoral research fellow at Harvard University. His research focuses on developing and applying statistical methodologies to answer questions in areas such as computational biology, cancer research and climate change.

A. H. Welsh is E. J. Hannan Professor of Statistics in the Research School of Finance, Actuarial Studies and Statistics at the Australian National University. His research focuses on statistical inference, statistical modelling, robustness, nonparametric and semiparametric methods, analysis of sample surveys and ecological monitoring.