Subseasonal Climate Prediction in the Western US using Bayesian Spatial Models

Vishwak Srinivasan[1] **Justin Khim**[3] **Arindam Banerjee**[2] **Pradeep Ravikumar**[1]

[1]Machine Learning Department, Carnegie Mellon University,
[2]Department of Computer Science, University of Illinois at Urbana-Champaign,
[3]Amazon

Abstract

Subseasonal climate forecasting is the task of predicting climate variables, such as temperature and precipitation, in a two-week to two-month time horizon. The primary predictors for such prediction problem are spatio-temporal satellite and ground measurements of a variety of climate variables in the atmosphere, ocean, and land, which however have rather limited predictive signal at the subseasonal time horizon. We propose a carefully constructed spatial hierarchical Bayesian regression model that makes use of the inherent spatial structure of the subseasonal climate prediction task. We use our Bayesian model to then derive decision-theoretically optimal point estimates with respect to various performance measures of interest to climate science. As we show, our approach handily improves on various off-the-shelf ML baselines. Since our method is based on a Bayesian framework, we are also able to quantify the uncertainty in our predictions, which is particularly crucial for difficult tasks such as the subseasonal prediction, where we expect any model to have considerable uncertainty at different test locations under different scenarios.

1 INTRODUCTION

Climate forecasts, which involve predictions of key climate variables such as temperature and precipitation, have immense utility to individuals, businesses, and government agencies [National Research Council, 2010]. These forecasts can be distinguished in terms of their prediction time horizons. *Weather forecasting* consists of predicting the day-to-day weather. These are largely based on numerical weather prediction (NWP) model typically based on dynamical models [Lorenc, 1986, Lorenz, 1996, Simmons and Hollingsworth, 2002] that can effectively forecast to about 10 days in advance [Bauer et al., 2015]. Since weather is primarily subject to atmospheric dynamics, and the atmosphere does not have substantial long term memory, this 10-day horizon is often regarded as the limit of predictability for numerical weather prediction models.

On the other hand, *climate forecasts* focus on predicting variables over several months (seasonal) to several decades (multi-decadal). Aggregated on such time frames, many of the day-to-day aberrations of the atmosphere are smoothed out, and consequently the coupling of the atmosphere with other components of the climate system, notably ocean, land, and sea ice, become more important. For example, at seasonal time scales, ocean variables often provide valuable predictive information because of the longer term memory of oceans. More generally, both simplified climate models as well as statistical models [Barnston et al., 2012] can be used quite effectively to produce seasonal forecasts.

Between weather and seasonal climate forecasts lies the regime of *subseasonal climate forecasting* (SSF). Here the goal is to predict at least two weeks to the future [White et al., 2017, DelSole and Banerjee, 2017, Totz et al., 2017, Raff et al., 2017, Cohen et al., 2019, Hwang et al., 2019, He et al., 2020]. In general, there are three main sources of predictability for climate forecasts: atmosphere, ocean, and the land, and the coupling between these climate variables play on important role; see Figure 1 in He et al. [2020]. What makes subseasonal prediction a far more difficult task than seasonal climate forecasting is that the shorter time horizon entails that the far more chaotic atmospheric noise is still an important factor, unlike for climate scale predictions. But on the other hand, it is also more difficult than weather forecasting because one needs to account for more than atmospheric climate variables given the longer time horizon.

Notwithstanding these structural bottlenecks to predictability, the subseasonal forecasting time frame is of significant practical importance, as has been noted by two National Academy of Science reports [National Research Council,

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:961–970.

2010, Ocean Studies Board and National Academies of Sciences, Engineering, and Medicine, 2016]. For instance, in agriculture, irrigation, pesticide, and fertilizer schedules must be implemented weeks in advance of an intended harvest [White et al., 2017]. We thus have the state of affairs that subseasonal predictions are crucially important for informed decision-making, and yet, subseasonal predictions are quite poor. There has thus been some burgeoning interest (as also recommended in the NAS reports cited above) in statistical machine learning based approaches for subseasonal forecasting, beyond just dynamical systems based climate models. One argument is that at longer lead times, the nonlinear atmospheric processes are too chaotic to be of use [Van den Dool, 2007]. Additionally, Cohen et al. [2019] argues that the proliferation of successful new statistical techniques could also be of use in climate forecasting. Indeed, in the recent subseasonal forecasting Rodeo, a climate prediction contest for the western US sponsored by NOAA and the US Bureau of Reclamation, simple yet thoughtful statistical models consistently outperformed NOAA's dynamical systems forecasts [Hwang et al., 2019].

To motivate our proposed approach, it will be instructive to discuss key facets of the subseasonal forecasting problem. Firstly, the data is noisy, high-dimensional, and highly spatially correlated, both in the covariate and response climate variables. The noise is due to chaotic nature of the atmosphere, making predictions weeks in advance difficult. The high dimensionality of the data is due to both the high spatial resolution of the climate variable measurements, as well as the size of the relevant regions. For instance, a large portion of the Pacific Ocean might include over 30,000 grid points for measurements of sea surface temperature, and even subsampling a smaller portion of the Pacific with a coarser grid can still lead to over 1,000 grid points. While a popular approach for such high-dimensional data is to impose sparsity, this might not be the best fit for the subseasonal prediction setting, since it is not clear if the target response could be adequately modeled using a few selected covariates. In contrast, the main structure in the data is its spatial smoothness, i.e., that nearby locations often have similar climate measurements. Thus, to effectively mitigate noise in the subseasonal regime, it might be preferable to carefully leverage spatial smoothness.

In this paper, we propose a simple and interpretable statistical machine learning approach, for subseasonal forecasting of temperature averaged over days 15 through 28 to the future on a $2^\circ \times 2^\circ$ grid over the western US, where we draw from approaches spatial statistics and econometrics [Gelfand et al., 2010, Cressie and Wikle, 2011, Banerjee et al., 2014], together with structural insights specific to the climate domain. Specifically, we use a spatial Bayesian hierarchical linear model to impose spatial structure on both the noise and the fitted regression coefficients. A key difference between our spatial Bayesian hierarchical linear model and other recent efforts in subseasonal climate forecasting Hwang et al. [2019], He et al. [2020] is that our modelling approach is Bayesian, and provides *probabilistic estimates* as compared to *point estimates*. There are two key advantages of our Bayesian approach. First, it allows us to obtain decision-theoretically optimal point predictions for common loss functions used in climate science, without having to fit a separate model for each loss function. Second, it allows us to quantify the uncertainty of our predictions, which is crucial in climate prediction in general.

In climate prediction, one usually seeks to improve on the forecast of the constant baseline known as *climatology*: the 30-year mean of a variable at a given location and day-of-the-year. We use the 30-year period from 1981 through 2010 to calculate the climatology and to fit our model, and we use 2011 through 2018 for evaluation. Two prediction metrics popular in climate science are mean-squared-error (MSE) skill and cosine similarity. As we discuss in more detail in our experiments, aggregating over all points, our models achieve skills of approximately 0.055 and 0.053 on the train and test sets. The constant baseline of climatology has a skill of zero, so the above is remarkably high for the subseasonal forecasting setting, and is also indicative of good generalization. We provide a detailed breakdown of our results from spatial and temporal standpoints, from which we observe that performance has considerable heterogeneity over both space (locations) and time (dates). Finally, our Bayesian spatial model allows us to examine and quantify the uncertainty of our predictions.

The remainder of the paper is organized as follows. In Section 2, we describe the data we use for our forecasting approach. In Section 3, we introduce our Bayesian spatial model, and in Section 4, we discuss how to derive decision-theoretically point predictions for varied loss functions used in climate science. In Section 5, we discuss our main results, and examine model fit, and compare to off-the-shelf modern ML baselines. And in Section 6, we discuss future directions for the subseasonal climate forecasting problem. Finally, our Appendices contain a number of additional details including further related work, specification of our data processing procedures, choice of hyperparameters of both our and baseline procedures, and more extensive comparisons with modern machine learning models.

2 DATA

We briefly review our data sources, and our process of standardization.

Our *target climate variable* is `tmp2m`, the air temperature measured at two meters above the ground in Celsius. We use data from the Climate Prediction Center (CPC) [Fan and Van den Dool, 2008], which is based on daily observations made since 1979 from the Global Historical Climatology

Network version 2 and the Climate Anomaly Monitoring System. The raw data is with respect to a $0.5° \times 0.5°$ latitude-longitude grid, which for computational considerations, we subsample to a $2° \times 2°$ grid. This subsampling process retains most of the spatial smoothness in the original data which is crucial to our method.

As *input climate variables*, we use sea surface temperature, sea level pressure, geopotential height, and relative humidity. The first climate variable, sea surface temperature (`sst`) is provided in degrees Celsius; with measurements from September 1981. Our second climate variable, sea level pressure (`slp`), is provided in pascals; with measurements from January 1981. Our last two climate variables are atmospheric: geopotential height at the 500 millibar level (`hgt500`) measured in meters; and relative humidity at the sigma level 995 measured in percentage (`rhum.sig995`). For all these variables, we use a $2° \times 2°$ latitude-longitude grid over oceans, subsampled from a $0.25° \times 0.25°$ grid of daily values [Reynolds et al., 2007]. Both the target and input climate variables are thus *spatio-temporal* in nature. We next discuss our standardization process.

First, due to our focus on the 2-4 week out subseasonal forecasting problem, we compute running averages of the data across a time period of 14 days for each location.

Second, as is standard in climate science, instead of predicting the raw temperature, we focus on and compute temperature *anomalies*, i.e., the deviation from the so-called *climatology*, which is the 30-year average temperature in Celsius for a given location and day of the year or two-week average ending on a given day of the year. Our reference climatology consists of the average over the years 1981 through 2010, as it is standard to compute climatology beginning with a year ending in 1.

Third, we standardize the anomalies using a simple z-score method, i.e., dividing the anomaly by its standard deviation, which gives rise to a *standardized anomaly* or a *z-score*. This is particularly crucial since climate variables have wildly different units and scaling: `slp` is measured in pascals with a range of approximately $9.0 \cdot 10^4$ Pa to $1.1 \cdot 10^5$ Pa, whereas `sst` is measured in Celsius with a range of approximately 4° C to 32° C. Given the predicted z-score from our models, we can then multiply with the climatology standard deviation, to then obtain our prediction of the anomaly itself (which with the climatology added, would yield the predicted temperature). We discuss data standardization in more detail in the Appendix.

3 METHODS

In this section, we introduce our Bayesian spatial model that is carefully constructed to leverage high level climate scientific domain insights, while drawing from spatial statistics and econometrics.

3.1 SETUP

We begin by setting up our notation. We let S denote the number of spatial locations for which we will be predicting. For each of these locations, we will be making predictions every 15 days; we let T denote the number of time periods that we will make predictions. Given the spatio-temporal climate covariates, we will be constructing predictors derived from PCA. Specifically, we obtain d predictors (which we might also term as input features) by computing the principal components of the z-scores of climate variables — `sst`, `slp`, `hgt500` and `rhum.sig995` — focusing on sea-based data. This is due to a climate scientific domain insight that a key source of predictors with "memory" sufficient to predict at subseasonal time scales are likely water-based.

Note that we obtain these PCs in a coupled fashion: we have 4 climate variables at S_{sea} locations over the sea, for a total of $4S_{\text{sea}}$ rows, from which we obtain the d covariates.

We collate these predictors into the design matrix $\mathbf{X}^{T \times d}$ as

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & \ddots & \vdots \\ x_{T,1} & \cdots & x_{T,d} \end{bmatrix}.$$

The target is multi-dimensional (we are predicting scalar temperature but at S locations). We collate the targets at all timesteps by the matrix $\mathbf{y} \in \mathbb{R}^{T \times S}$ given by

$$\mathbf{y} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,S} \\ \vdots & \ddots & \vdots \\ y_{T,1} & \cdots & y_{T,S} \end{bmatrix}.$$

We denote $\mathbf{y}_t = (y_{t,1}, \ldots, y_{t,S})$ as the t^{th} row of $\mathbf{y}$, which is all of the target values across spatial locations for a given time t.

3.2 SPATIAL MODEL

We next discuss our Bayesian spatial regression model.

Before we specify the Bayesian setup, consider the multiple linear regression model for predicting the target $\mathbf{y}_t$ given the covariates $\mathbf{X}_t$:

$$\mathbf{y}_t = \beta^T \mathbf{X}_t + \epsilon_t, \tag{1}$$

for all $t \in \{1, \ldots, T\}$, where $\epsilon_t \in \mathbb{R}^S$ is a zero-mean noise variable, and the multiple regression coefficients are collated into the matrix $\beta \in \mathbb{R}^{d \times S}$ as

$$\beta = \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,S} \\ \vdots & \ddots & \vdots \\ \beta_{d,1} & \cdots & \beta_{d,S} \end{bmatrix}.$$

In other words, β is the matrix of row vectors, each of which are the collection of the coefficients for a covariate across spatial locations. Note that this models observations at different times (where each consecutive pair of timestamps are 15 days apart) as being independent.

There are two caveats to simply fitting a vanilla multiple linear regression:

1. The predictions need not be spatially smooth. As we discuss below, we ensure this by enforcing spatial smoothness of regression coefficients via a carefully constructed prior
2. The noise vector term ϵ_t need not be independent across all spatial locations. We allow for spatial correlation of the noise term.

Before specifying these spatial smoothness and dependence modifications, we first set up some notation. We define a *spatial graph* over the target region, with vertices of this graph as latitude-longitude pairs, and edges between vertices with locations *adjacent* to each other in the latitude-longitude grid under consideration. We define the notion of *adjacency* as follows.

Definition 1 (Adjacency)**.** *Two locations* $u = (u_{lat}, u_{lon})$ *and* $v = (v_{lat}, v_{lon})$ *are spatially adjacent with respect to a set of locations* $\mathcal{S}$ *if*

$$\mathsf{dist}(u, v) := \|u - v\|_2 = \mathsf{res}(\mathcal{S}),$$

where $\mathsf{res}(\mathcal{S}) = \min_{x,y \in \mathcal{S}; x \neq y} \mathsf{dist}(x, y)$ *is the resolution of* $\mathcal{S}$*.*

Example: Consider any non-boundary location u in the $2^\circ \times 2^\circ$ latitude-longitude grid of points for `tmp2m`. Then, its adjacent points are simply the grid points 2° to the north, south, east, and west of u.

The adjacency matrix $A \in \{0, 1\}^{S \times S}$ of this spatial graph is a matrix with $A_{u,v} = 1$ if u and v are distinct and adjacent and zero otherwise. Let D in $\mathbb{R}^{S \times S}$ denote the diagonal degree matrix where $D_{s,s}$ is the degree of s in $\mathcal{S}$. Then, the Laplacian matrix $L \in \mathbb{R}^{S \times S}$ is given by $L = D - A$.

Given this notation, we define the following parameterized family of spatial covariance matrices

$$\Sigma(\rho) := (\rho L + (1 - \rho) I_S)^{-1}, \tag{2}$$

parameterized by ρ in $(0, 1)$. Note that we often drop the ρ dependence on the left hand side when conditioning on ρ. This shrinkage model can be compared to Tikhonov regularization studied in Belkin et al. [2004].

While this construction might not seem as intuitive at first sight, its importance will become clearer once we specify our spatial prior over the regression coefficients, as follows:

$$\beta_j \mid \beta_{0_j}, \tau_j^2 \sim \mathcal{N}(\beta_{0_j}, \tau_j^2 \Sigma), \tag{3}$$

for each $j \in \{1, \dots, d\}$, where β_j is the j^{th} row of β defined earlier. Note that the spatial dependence in the regression coefficients arises due to our spatial covariance Σ, which is scaled by τ_j^2. Furthermore, if we considered the regression parameter β to be a vector of length $d \cdot S$, then the model can be concisely expressed as

$$\mathrm{vec}(\beta) \mid \mathrm{vec}(\beta_0), \tau^2 \sim \mathcal{N}(\mathrm{vec}(\beta_0), \Omega_0^{-1}) \tag{4}$$

where the prior mean is

$$\mathrm{vec}(\beta_0) = (\underbrace{\beta_{0,1}, \dots, \beta_{0,1}}_{S \text{ times}}, \underbrace{\beta_{0,2}, \dots, \beta_{0,2}}_{S \text{ times}}, \dots, \beta_{0,d})$$

and the prior covariance is the block diagonal matrix

$$\Omega_0 = \begin{bmatrix} \frac{1}{\tau_1^2}\Sigma^{-1} & & \\ & \ddots & \\ & & \frac{1}{\tau_d^2}\Sigma^{-1} \end{bmatrix} = \mathrm{diag}(1/\tau^2) \otimes \Sigma^{-1}.$$

Here τ^2 is the length d vector of τ_j^2s specified in (3) and $A \otimes B$ denotes the Kronecker product between A and B.

We now have the notation necessary to detail the motivation behind our spatial covariance matrix construction in (2).

Let $\mathrm{vec}(\beta)_{-j,-s}$ denote the vector in $\mathbb{R}^{d \cdot S - 1}$ obtained by removing $\beta_{j,s}$ from $\mathrm{vec}(\beta)$.

Proposition 1. *Suppose the regression coefficients have the prior as specified in* (3)*. Then, the conditional prior distribution of* $\beta_{j,s}$ *is*

$$\begin{aligned} &\beta_{j,s} \mid \mathrm{vec}(\beta)_{-j,-s},\ \mathrm{vec}(\beta_0),\ \tau^2 \\ &\quad \sim \mathcal{N}\left(\frac{\rho \sum_{s' \neq s} A_{s,s'} \beta_{j,s'} + (1 - \rho)\beta_{0,j}}{\rho \sum_{s' \neq s} A_{s,s'} + (1 - \rho)}, \right. \\ &\qquad\qquad \left. \frac{\tau_j^2}{\rho \sum_{s' \neq s} A_{s,s'} + (1 - \rho)} \right). \end{aligned} \tag{5}$$

Proof. The equivalence between (4) and (5) follows from an application of Brook's lemma [Banerjee et al., 2014, Brook, 1964]. □

The proposition thus motivates our spatial covariance matrix construction in (2). When using this covariance matrix within the prior in (3) on the regression coefficients, the prior could alternatively be considered as a spatial autoregressive nature of the model, respecting the adjacency graph as defined earlier.

Lastly, we define the prior over the noise terms as follows:

$$\epsilon_t \mid \rho, \tau_\epsilon^2 \sim \mathcal{N}(\mathbf{0}, \tau_\epsilon^2 \Sigma)$$

to be spatially-correlated noise. We further assign a prior for τ_ϵ^2 as

$$\tau_\epsilon^2 \mid a_\epsilon, b_\epsilon \sim \text{Inverse-Gamma}(a_\epsilon, b_\epsilon)$$

We choose the hyperparameters a_ϵ, b_ϵ in an empirical Bayes manner, as discussed in the Appendix.

The priors over the regression coefficients and noise terms are in turn specified by various hyperparameters. We impose additional priors over these to build a hierarchical Bayesian model. For the means β_{0_j}, we set

$$\beta_{0_j} \mid \lambda \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\lambda ST} I\right). \tag{6}$$

Note that λ plays the role of a regularization hyperparameter; higher λ indicates stronger regularization and vice versa. We then place a prior over λ as

$$\lambda \mid a_\lambda, b_\lambda \sim \mathrm{Gamma}(a_\lambda, b_\lambda).$$

where (a_λ, b_λ) are chosen to be $(10, 10)$. Next, for the prior variances, we set

$$\tau_j^2 \sim \text{Inverse-Gamma}(a_j, b_j), \tag{7}$$

and the hyperparameters (a_j, b_j) are chosen in an empirical Bayes manner. For further details, see the Appendix. Next, we need to place a prior on ρ, which controls the spatial covariance. We select the prior

$$\rho \sim \mathrm{Beta}(a_\rho, b_\rho), \tag{8}$$

where (a_ρ, b_ρ) are chosen to be $(8, 2)$.

Our model is implemented using NumPyro [Phan et al., 2019]. We use the No-U-Turn-Sampler (NUTS) [Hoffman and Gelman, 2014] for inference based on Hamiltonian Monte Carlo (HMC). In particular, NUTS provides an efficient way of running HMC, and enables faster convergence to the target posterior distributions, and we verify convergence using the Gelman-Rubin statistic which is also computed in NumPyro.

Computationally, one difficulty is that the number of parameters dS is quite large. In our computations, $d = 20$ and $S = 2004$ leading to $dS = 40080$. Thus, a smaller S leads to much faster computations, which motivated our subsampling of the target climate variable `tmp2m` to over a 2° × 2° grid. With the original resolution, we had $2,004$ spatial grid points, whereas with the subsampling this reduced to 132, while nonetheless retaining most of the spatial smoothness discussed earlier. Despite this, NUTS struggles to scale, and this is due to the number of unobserved variables in the model. Therefore, one simplification we make before running our model is to *collapse* the prior on the mean of β. Due to conjugacy of this prior (Normal-Normal), we marginalize out effect of β_0, and consequently leads to our model running faster.

4 FROM BAYESIAN MODELS TO DECISION-THEORETIC PREDICTION

Our Bayesian spatial regression models provide the distribution of the target climate variable (conditioned on the covariates) at any location and time. Here, we discuss how to obtain *point estimates* given these distributions and the specific decision-theoretic loss functions used in climate science: squared-error skill and cosine similarity. This is a key advantage of Bayesian machinery: we can easily derive *skillful* point predictions given any new loss function of interest, without any laborious retraining of models.

4.1 LOSS FUNCTIONS / PREDICTION METRICS

Before we present the loss functions / prediction metrics, we first set up some notation. We let $y_{s,t}^{\mathrm{clim}}$ denote the climatology of temperature i.e., the 30 year average for `tmp2m` at location s and time t as described in Section 2. Let $y_{s,t}$ denote the true measurement made at location s and time t, and $\widehat{y}_{s,t}$ denote the prediction made at location s and time t. Consequently, we define the following relative values

$$\begin{aligned} a_{s,t} &= y_{s,t} - y_{s,t}^{\mathrm{clim}} \\ \widehat{a}_{s,t} &= \widehat{y}_{s,t} - y_{s,t}^{\mathrm{clim}}. \end{aligned}$$

The key prediction metric used in climate science is *skill* [Van den Dool, 2007]. This metric specifically measures the relative performance of our prediction in comparison to the climatology. This is given by

$$\begin{aligned} \mathrm{skill}(\mathcal{S}', \mathcal{T}') &= 1 - \frac{\sum_{s\in\mathcal{S}'}\sum_{t\in\mathcal{T}'}(y_{s,t} - \widehat{y}_{s,t})^2}{\sum_{s\in\mathcal{S}'}\sum_{t\in\mathcal{T}'}(y_{s,t} - y_s^{\mathrm{clim}})^2} \\ &= 1 - \frac{\sum_{s\in\mathcal{S}'}\sum_{t\in\mathcal{T}'}(a_{s,t} - \widehat{a}_{s,t})^2}{\sum_{s\in\mathcal{S}'}\sum_{t\in\mathcal{T}'}(a_{s,t})^2}. \end{aligned}$$

Here, the two arguments $\mathcal{S}', \mathcal{T}'$ are sets of spatial locations, and sets of specific times respectively. In-sample, this might seem the same as the usual statistical notion of R^2; however, here, the sample mean that would appear in the denominator for R^2 is replaced by the climatology instead.

Another metric used in climate science is *cosine similarity*; for instance, this was used in the Subseasonal Forecast Rodeo and related work [Raff et al., 2017, Hwang et al., 2019]. In particular, the contest considers spatial similarity

$$\text{cos-sim}(\mathcal{S}', \mathcal{T}') = \frac{\langle a_{\mathcal{S}',\mathcal{T}'}, \widehat{a}_{\mathcal{S}',\mathcal{T}'}\rangle}{\|a_{\mathcal{S}',\mathcal{T}'}\|_F \|\widehat{a}_{\mathcal{S}',\mathcal{T}'}\|_F}.$$

Note that for the Rodeo contest, cosine similarity is computed spatially, i.e., $\mathcal{T}' = \{t\}$ consists of a single date, and these are averaged or otherwise compared over time. Alternatively, we consider cosine similarity across space and time, i.e., $\text{cos-sim}(\mathcal{S}, \mathcal{T})$. However, the difference between averaged spatial cosine similarity and total cosine similarity i.e., with $\mathcal{T}' = \{1, \ldots, T\}$, as we use, tends to be small. We briefly note here that cosine similarity is unusual or less than ideal from a statistical perspective, since it ignores the scale of predictions and only considers the cosine of the angle between the vector of predictions from various grid points; it is nonetheless popular in climate science.

Method	Train skill	Test skill	Train cos-sim	Test cos-sim
Spatial-regression	$5.48 \cdot 10^{-2}$	$\mathbf{5.32 \cdot 10^{-2}}$	$2.34 \cdot 10^{-1}$	$\mathbf{2.31 \cdot 10^{-1}}$
XGBoost	$\mathbf{1.52 \cdot 10^{-1}}$	$4.39 \cdot 10^{-2}$	$\mathbf{4.23 \cdot 10^{-1}}$	$2.13 \cdot 10^{-1}$
Multi-task Lasso	$5.53 \cdot 10^{-2}$	$4.98 \cdot 10^{-2}$	$2.36 \cdot 10^{-1}$	$2.23 \cdot 10^{-1}$
Neural Network	$5.54 \cdot 10^{-2}$	$4.95 \cdot 10^{-2}$	$2.35 \cdot 10^{-1}$	$2.24 \cdot 10^{-1}$

Table 1: Overall metrics on the train and test sets; higher is better. Note that the spatial regression approach performs the best with respect to both metrics on the test set.

4.2 BAYES OPTIMAL POINT PREDICTIONS

Given the loss functions above, we next discuss computing decision theoretically optimal point estimates with respect to our Bayesian regression model.

Let us focus on predictions at particular time t. For any of our loss functions $\mathcal{L}(\mathcal{S}', \{t\})$, let us no longer suppress the dependence on the true anomalies $\mathbf{a}$ and the predicted anomalies $\widehat{\mathbf{a}}$, so that we use $\mathcal{L}(\mathcal{S}', \{t\})[\mathbf{a}, \widehat{\mathbf{a}}]$ in the development below.

Given our Bayesian model for the conditional distribution of the anomalies, $P(\mathbf{a}_t|\mathbf{x}_t)$, we can compute the conditional expected loss

$$\mathcal{L}_P(\widehat{\mathbf{a}}_t) = \mathbb{E}_{\mathbf{a}\sim P(\cdot|\mathbf{x}_t)}\mathcal{L}(\mathcal{S}', \{t\})[\mathbf{a}, \widehat{\mathbf{a}}].$$

The decision-theoretically optimal point prediction for the Bayesian model $P(\mathbf{a}_t|\mathbf{x}_t)$ with respect to the loss function $\mathcal{L}(\cdot,\cdot)$ is then given by $\widehat{\mathbf{a}}_t \in \operatorname{argmin}_{\widehat{\mathbf{a}}_t} \mathcal{L}_P(\widehat{\mathbf{a}}_t)$.

We next derive these decision theoretically optimal point predictions for the two loss functions we had discussed earlier. In the sequel, we focus on a specific time t, and condition on covariates $\mathbf{x}_t \in \mathbb{R}^d$. We begin by noting that

$$\text{skill}(\mathcal{S}', \{t\}) \propto -\sum_{s\in\mathcal{S}'}(a_{s,t} - \widehat{a}_{s,t})^2.$$

Therefore, maximizing the *skill* is equivalent to minimizing the squared error loss. It then follows that the decision-theoretically optimal point prediction is given as:

$$\widehat{a}_t = \operatorname*{argmin}_c \mathbb{E}_{a'\sim P(a|\mathbf{x}_t)} \|a' - c\|^2 = \mathbb{E}_{a'\sim P(a|\mathbf{x}_t)}[a'].$$

Drawing N samples $a'_1, \ldots, a'_N$ from the posterior distribution, we can compute the usual Monte-Carlo approximation

$$\widehat{a}_t = \frac{1}{N}\sum_{i=1}^{N} a'_i. \quad (9)$$

Analogously, we can also derive decision-theoretic optimal point predictions that maximize cosine similarity as

$$\widehat{a}_t = \operatorname*{argmax}_c \mathbb{E}_{a'\sim P(a|\mathbf{x}_t)} \frac{\langle a', c\rangle}{\|a'\|\|c\|}$$

$$= \operatorname*{argmax}_{c:\|c\|\le 1} \left\langle \mathbb{E}_{a'\sim P(a|\mathbf{x}_t)}\left[\frac{a'}{\|a'\|}\right], c\right\rangle = \mathbb{E}_{a'\sim P(a|\mathbf{x}_t)}\left[\frac{a'}{\|a'\|}\right]$$

As done earlier, drawing N samples $a'_1, \ldots, a'_N$ from the posterior distribution, we obtain the Monte-Carlo estimate

$$\widehat{a}_t = \frac{1}{N}\sum_{i=1}^{N}\frac{a'_i}{\|a'_i\|}.$$

Due to our focus on skill as the primary metric, we use (9) as our point predictions made by our model.

5 RESULTS

We start by providing results on the accuracy of our decision-theoretically optimal point prediction with respect to the squared error loss given our Bayesian spatial model, comparing to standard baselines. We also quantifying the uncertainty of our predictions via prediction coverage, and posterior intervals at the 95% confidence level. In all our experiments, we choose the number of principal component based covariates $d = 20$. Additionally, we predict standardized anomalies, which we rescale back to obtain the unstandardized predictions, as discussed earlier, and which we also detail further in the Appendix.

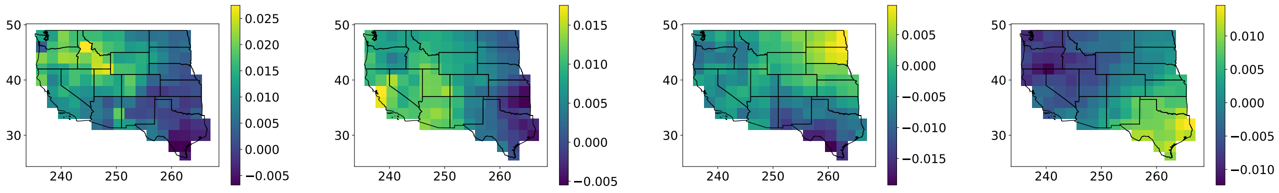
5.1 POINT PREDICTIONS

We first compare the quality of our prediction to various machine learning (ML) baselines using metrics defined earlier. For all these ML baselines, we use the same PCA derived covariates. The methods we include are (a) XGBoost [Chen and Guestrin, 2016], (b) Multi-task Lasso [Zhang et al., 2006] and (c) a multi-layer nonlinear neural network [Goodfellow et al., 2016]. More details regarding the hyperparameters of these baselines are specified in the Appendix.

We provide these point prediction evaluations in Table 1. As it shows, our Bayesian spatial model has markedly better performance on the test sets with respect to both skill and cosine similarity metrics. Interestingly, XGBoost has better training performance, but worse test performance than other approaches, which indicates that it does not generalize well. This difficulty to generalize has been noted by earlier work

Year	Spatial-regression	XGBoost	Multi-task Lasso	Neural Network
2011	$1.19 \cdot 10^{-1}$	$-4.32 \cdot 10^{-2}$	$\mathbf{1.28 \cdot 10^{-1}}$	$1.27 \cdot 10^{-1}$
2012	$1.26 \cdot 10^{-1}$	$\mathbf{1.84 \cdot 10^{-1}}$	$1.15 \cdot 10^{-1}$	$1.23 \cdot 10^{-1}$
2013	$-7.60 \cdot 10^{-2}$	$-1.04 \cdot 10^{-1}$	$\mathbf{-6.78 \cdot 10^{-2}}$	$-7.90 \cdot 10^{-2}$
2014	$\mathbf{3.99 \cdot 10^{-2}}$	$2.66 \cdot 10^{-2}$	$3.71 \cdot 10^{-2}$	$3.66 \cdot 10^{-2}$
2015	$-2.26 \cdot 10^{-2}$	$\mathbf{1.19 \cdot 10^{-1}}$	$-3.42 \cdot 10^{-2}$	$-3.51 \cdot 10^{-2}$
2016	$\mathbf{1.52 \cdot 10^{-1}}$	$1.46 \cdot 10^{-1}$	$1.44 \cdot 10^{-1}$	$1.51 \cdot 10^{-1}$
2017	$5.59 \cdot 10^{-2}$	$\mathbf{7.55 \cdot 10^{-2}}$	$5.05 \cdot 10^{-2}$	$4.98 \cdot 10^{-2}$
2018	$\mathbf{5.75 \cdot 10^{-2}}$	$-3.16 \cdot 10^{-2}$	$5.68 \cdot 10^{-2}$	$5.36 \cdot 10^{-2}$

Table 2: Variation of skill annually on the test set. Note the considerable heterogeneity across years.



Covariate (1)

Covariate (2)

Covariate (3)

Covariate (4)

Figure 2: The posterior means of regression coefficients by spatial location for 4 randomly chosen covariates.

as a key facet of subseasonal forecasting problem, indicating that the Bayesian machinery provides crucial adaptive regularization that allows for resistance to overfitting.

To provide further insights into these results, we break down the test performance by year in Table 2. The most important takeaway is that there is considerable heterogeneity in performance by year. We see that for a majority of the models, 2013 and 2015 appear to be difficult years to predict, since the skills are the lowest then, whereas XGBoost seems to perform surprisingly well in 2015, while performing suboptimally in the years where other models perform well. The difficulties in 2013 and 2015 could be due to the extreme winter and the El Nĩno event that occurred in the respective years. In the Appendix, we provide the annual variation in cosine-similarity as well.

In addition to the temporal breakdown of the performance above, we also provide a spatial breakdown, by plotting the average skill for any location in Figure 1. The model can be seen to perform well in the Pacific Northwest and Texas, while performing relatively poorly in California and the Arizona-New Mexico border. We show the spatial variation in cosine similarity for our model, as well as the spatial variation for the other ML baselines in the Appendix.

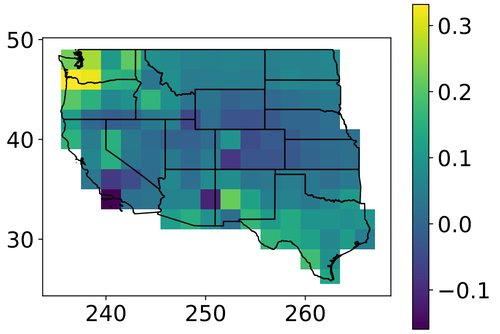


Figure 1: Variation of skill by location. Note the variance across locations, especially those closer to the coast.

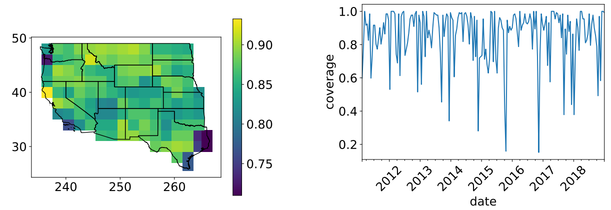
5.2 INSPECTING THE REGRESSION COEFFICIENTS

A crucial advantage of our linear spatial model is its interpretability. In particular, we can gain an understanding of our fitted model by inspecting the regression coefficients for each of the $d = 20$ covariates. In Figure 2, we plot the posterior means of regression coefficients at each spatial location corresponding to 4 randomly chosen covariates; we provide the rest in the Appendix.

As can be seen in the figure, some interesting spatial patterns emerge. Consistent with our modeling, we observe smoothness in the regression coefficients. Note that for the 1^{st} covariate, we see that the strongest coefficients are positive, and in the states of Idaho, and disperse around that state. On the other hand, for the 2^{nd} covariate, we see the strongest coefficients in California, Nevada, Arizona and Utah, and the effect gradually weakens towards the east. The 3^{rd} and 4^{th} covariates have their strongest coefficients that are more eastward, while there is a negative effect of equal magnitude on the west around the states of Washington and Oregon.

5.3 QUANTIFYING PREDICTION UNCERTAINTY

A crucial advantage of our Bayesian machinery is that it allows for easy quantification of the uncertainty of our predictions.

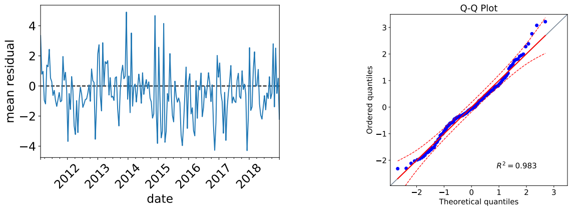


(a) (b)

Figure 3: Spatial and temporal coverage for our prediction intervals.

We start with the coverage of high-confidence posterior prediction intervals. First, we construct 95% posterior prediction intervals for each time and location: these are intervals in the target response domain that have more than 95% probability with respect to our Bayesian model posterior. Next, we computed the fraction of train and test points that fall in this interval: the train coverage was 0.89, while the test coverage was 0.86. While large, this is lower than a typically expected 0.95 coverage level in parametric Bayesian analyses. This is due to the difficulty of the subseasonal forecasting. From a statistical perspective, this is likely due to the fact that distributions of temperature are known to have relatively heavy tails; so some *undercoverage* would be expected.

To provide further insights into the posterior prediction coverage as computed above, we breakdowns of the coverage with respect to both space and time in Figure 3. From Figure 3(a), we see that coverage is very spatially heterogeneous. However, it is generally better in the north and poorer down south. Additionally, there are a few spots of exceedingly poor coverage in Washington state, where perhaps mountains or rivers have important effects. We also observe this in state of Texas near the border to the east, where perhaps the Gulf of Mexico has a more prominent effect in predicting the temperature. With respect to temporal variation, from Figure 3(b), we observe that coverage can be quite temporally heterogeneous as well. In general, it appears that poor coverage dates tend to be in the winter months, although there are days with poor coverage in the summer of 2015.



(a) (b)

Figure 4: Variation of mean residuals temporally, and normal quantile-quantile for the mean residual across the Western US.

One hypothesis for the heterogeneity above could be the persistent heteroscedasticity, which we consider in our examination on the noise. To get additional insights on the noise distribution, we next analyze the residuals with respect to our point predictions in greater detail, as shown in Figure 4. In panel (a), we plot the mean residuals over time, from which we can immediately observe the heteroscedasticity over time. Indeed, it appears that residuals are generally larger in the winter months. A second observation is possibly a weak trend: while at any point it seems the signs of the residuals may be hard to predict, in later years, the average residual appears to be negative. This implies the predictions are consistently high for this time period. In panel (b), we provide a quantile-quantile plot for the residuals averaged across the Western US, together with an envelope at the 95% level. We notice that the mean residuals over the Western US are approximately normal; one explanation for which is that even with spatial dependence, the central limit theorem applies. This indicates that our normal distributional assumptions on the noise was a reasonable fit to the data.

6 DISCUSSION

Our paper presents a Bayesian spatial model for subseasonal climate forecasting in the western US based on exploiting spatial smoothness in the data. We next list some possible avenues for improvements and future work. First, our modeling of the climate dynamics could be improved, by going beyond the features we use to incorporate more nuanced features such as eddies and seasonal currents, as well as other climate covariates, such as sea ice concentration, soil moisture or polar pressure anomalies. Second, from a modeling perspective, it might be helpful to build a two-staged model, where we identify locations where the standardized climate anomalies are not normal distributed and build a more complex model for such locations. Finally, it would be of interest to extend our subseasonal forecasts from the US to other parts of the world, assessing their ability to predict droughts that commonly lead to water shortages, or floods that lead to infectious disease outbreaks through mosquitoes, among others.

Acknowledgements

VS, JK and PR acknowledge the support of NSF via OAC-1934584. AB acknowledges the support of NSF via OAC-1934634, IIS-1908104, and a C3.ai grant. VS also thanks Neeraj Pradhan, Fritz Obermeyer and Du Phan for help with NumPyro.

References

Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC press, 2014.

Anthony G Barnston, Michael K Tippett, Michelle L L'Heureux, Shuhua Li, and David G DeWitt. Skill of real-time seasonal enso model predictions during 2002–11: Is our capability increasing? *Bulletin of the American Meteorological Society*, 93(5):631–651, 2012.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525 (7567):47–55, 2015.

Mikhail Belkin, Irina Matveeva, and Partha Niyogi. Regularization and semi-supervised learning on large graphs. In *International Conference on Computational Learning Theory*, pages 624–638. Springer, 2004.

D Brook. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, 51 (3/4):481–483, 1964.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Judah Cohen, Dim Coumou, Jessica Hwang, Lester Mackey, Paulo Orenstein, Sonja Totz, and Eli Tziperman. S2S reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *Wiley Interdisciplinary Reviews: Climate Change*, 10(2), 2019.

Noel Cressie and Christopher K Wikle. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, 2011.

Timothy DelSole and Arindam Banerjee. Statistical seasonal prediction based on regularized regression. *Journal of Climate*, 30(4):1345–1361, 2017.

Yun Fan and Huug Van den Dool. A global monthly land surface air temperature analysis for 1948–present. *Journal of Geophysical Research: Atmospheres*, 113(D1), 2008.

Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of Spatial Statistics*. CRC press, 2010.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Sijie He, Xinyan Li, Timothy DelSole, Pradeep Ravikumar, and Arindam Banerjee. Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances. *arXiv preprint arXiv:2006.07972*, 2020.

Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Jessica Hwang, Paulo Orenstein, Judah Cohen, Karl Pfeiffer, and Lester Mackey. Improving subseasonal forecasting in the western us with machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2325–2335, 2019.

Andrew C Lorenc. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112(474):1177–1194, 1986.

Edward N Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1, 1996.

National Research Council. *Assessment of intraseasonal to interannual climate prediction and predictability*. National Academies Press, 2010.

Ocean Studies Board and National Academies of Sciences, Engineering, and Medicine. *Next generation earth system prediction: strategies for subseasonal to seasonal forecasts*. National Academies Press, 2016.

Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

David Raff, Kenneth Nowak, Robert Cifelli, Levi D Brekke, and Robert Stabler Webb. Sub-seasonal climate forecast rodeo. In *AGU Fall Meeting*, 2017.

Richard W Reynolds, Thomas M Smith, Chunying Liu, Dudley B Chelton, Kenneth S Casey, and Michael G Schlax. Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate*, 20(22):5473–5496, 2007.

Adrian J Simmons and Anthony Hollingsworth. Some aspects of the improvement in skill of numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 128(580): 647–677, 2002.

Sonja Totz, Eli Tziperman, Dim Coumou, Karl Pfeiffer, and Judah Cohen. Winter precipitation forecast in the european and mediterranean regions using cluster analysis. *Geophysical Research Letters*, 44(24):12–418, 2017.

Huug Van den Dool. *Empirical methods in short-term climate prediction*. Oxford University Press, 2007.

Christopher J White, Henrik Carlsen, Andrew W Robertson, Richard JT Klein, Jeffrey K Lazo, Arun Kumar, Frederic Vitart, Erin Coughlan de Perez, Andrea J Ray, Virginia Murray, et al. Potential applications of subseasonal-to-seasonal (s2s) predictions. *Meteorological applications*, 24(3):315–325, 2017.

Jian Zhang, Zoubin Ghahramani, and Larry Wasserman. A probabilistic framework for multi-task learning, 2006.