Escola de Camins
Escola Tècnica Superior d'Enginyeria de Camins, Canals i Ports
UPC BARCELONATECH

Space-time Bayesian Hierarchical Model for paleoclimate reconstruction

TREBALL FINAL DE MÀSTER

Treball realitzat per:
Francisco Javier Gual Navarro

Dirigit per:
Balaji Rajagopalan
Irene Arias

Màster en:
Enginyeria de camins, canals i ports.

Barcelona, 28 de setembre de 2022

Departament d'enginyeria civil i ambiental

Treball Final de Master

A Space-Time Bayesian Hierarchical Model for Climate Anomalies Over a Large Domain : Application to Paleoclimate Reconstruction of Sea Surface Temperatures over Equatorial Pacific Ocean

Francisco Javier Gual Navarro
Advisor: Balaji Rajagopalan
UPC supervisor: Irene Arias

September 28, 2022

Abstract

From 10 to 2 ka, the entire equatorial Pacific warms, but at a faster rate in the east than in the west. This pattern is broadly consistent with previous inferences of reduced El Niño-Southern Oscillation variability associated with a "La Niña-like" state during the early to middle Holocene. Reduced space methods for annual sea surface temperature (SST) reconstruction over the equatorial Pacific have been proposed but their ability to quantify uncertainties is limited.

Motivated by this gap, in this study, a spatial and temporal Bayesian hierarchical model (BHM) is proposed to reconstruct annual SST over the equatorial pacific [10^{o}S to 10^{o}N and 100^{o}E to 75^{o}W (285^{o}E)]. First, dominant modes of variability of the contemporary (1854 – 2014) SST field over the locations of proxy (i.e., sediment core) data are obtained via Principal Component Analysis. The SST over the entire domain is modeled as a Multivariate Normal distribution in the data layer of the BHM. The mean and standard deviation are modeled to vary spatially as a function of the dominant modes in the process layer. A second process layer involves using Gaussian Kernels to model the spatial covariance. Suitable priors are used, and via Markov Chain Monte Carlo (MCMC), the posterior distribution of the parameters and, consequently, that of the SST fields are obtained for any desired year. The model is developed and validated for the contemporary period and subsequently applied for reconstructing SST fields during the Holocene. Reconstructions are made over several years during the Holocene, and El Niño Southern Oscillation (ENSO) indices and their respective uncertainties are obtained. The proposed model is one of the first attempts at reconstructing SSTs over the entire equatorial Pacific domain jointly with robust quantification of uncertainties.

A second model, less complex, is proposed. In this case a semi-Bayesian model, working both with the Principal Components of both the data in the domain and the data over the locations of proxy. A brief comparison is shown between both results.

Keywords: statistics, Bayesian modeling, El Niño, La Niña, Pacific, uncertainty.

Contents

1 Introduction

The equatorial Pacific has a big influence in some of the planet climate patterns. El Niño and la Niña related events, i.e., warming and cooling of the sea surface temperatures of the eastern equatorial Pacific, have direct impacts over rainfall, temperature, and winds [Delage and Power, 2020, NOA, UN2].

Understanding previous climate patterns in this region has been in the focus of several research studies, and different models have been proposed. One of the models presented by Gill et al. [2016], is one of the closest references to check the results. Most of them, including the one just mentioned, suggest that this Pacific region has been warming since the early Holocene, at different paces throughout the years. Also, some of them show that between 6 ka and 4 ka, the eastern equatorial Pacific had a smaller cooling, changing the trend of the previous Holocene years.

Core data from proxies located at the coasts of the equatorial Pacific, and recent findings of new data in the central Pacific store information of previous sea surface temperatures over the years. The relation between these points and the El Niño events can be found and modeled to understand what the previous eras looked like and how those climates impacted the rest of the world, thus, how they impacted the civilizations at that time.

However, paleoclimatic proxy data used are intrinsically point measurements, and therefore associating one paleoclimate time series with a process that reaches over a huge area takes a risk. Moreover, because of dissolution of carbonate tests and low sedimentation rates at great water depths [Gill et al., 2016], useful cores for SST proxy records are typically limited to regions of relatively shallow depth, and therefore near coasts or high mountains in the ocean, from aseismic ridges or over young oceanic crust. As a result, the absence of more long-term records from most of the areas in the equatorial Pacific makes it difficult to draw conclusions about large-scale spatial patterns of SSTs over paleoclimatic timescales. The records that exist along the ocean margins and ridges are often irregularly sampled (spatially and temporally, due to some uncertainty in both dimensions), and different proxies from the same core can sometimes yield contradictory inferences.

The results of a spatial-Bayesian reconstruction approach based on principal component predictors is developed and presented, to reconstruct equatorial Pacific mean SSTs over the Holocene. The method was applied to all available equatorial Pacific SST records from two common paleothermometric proxies: planktonic foraminiferal Mg/Ca and alkenone unsaturation [Gill et al., 2016].

No other studies presented such an uncertainty as this one, showing the posterior distributions of the objective variables in the area of study. However, similar approaches have been presented in other areas and fields [Ossandón et al., 2021, Tingley and Huybers, 2010], but not with the same structure presented in this work. The model presented uses probability to represent all uncertainties within the model, i.e., the parameters and output, both in space and time, which employs some knowledge in the parameters that model the information.

1.1 Objectives

The objectives of this work are presenting a spatial reconstruction of the climate at the equatorial pacific, working with a few data points as predictors. The approach allows to

reconstruct the climate in past periods of times, in these study, the paleo data used are proxy cores from the Holocene.

Another of the goals of this work is seeing if the semi-Bayesian approach, much simpler might be close to the full Bayesian model results. Sometimes simpler approaches give almost the same results, and more complex models might just add extra degrees of freedom in the parameter fitting.

Lastly, uncertainty in variables such as temperatures and climate is very important and other deterministic models have not been able to provide with statistical information of them in some of the reconstructions presented [Gill et al., 2016, 2017]. The Bayesian framework provides both, a model that allows for reconstructing temperatures and uncertainty of these variables. Both of them are studied in this work.

These objectives are presented along the work in the different sections, also with some lines of work for future works tin which this can be extended in the future.

1.2 State of the art

The models presented are one of the first attempts at reconstructing SSTs over the entire equatorial Pacific domain jointly with robust quantification of uncertainties. It is also one of the first space and time Hierarchical models, due to the introduction of both space and time predictors.

The Bayesian and semi Bayesian models presented in this work are based in the first principles of statistics, mainly in the Bayes' theorem.

The Bayes' theorem describes the probability of an event, based on prior knowledge of circumstances or conditions that may be related to the event.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{1}$$

where A and B are events and the event B has a probability strictly greater than zero.

- $P(A)$ is the probability of the event A without any given conditions.
- $P(B)$ is the probability of the event B without any given conditions.
- $P(A \mid B)$ is the conditional probability: the probability of event A occurring given that B is true. It is also called the posterior probability of A given B.
- $P(B \mid A)$ is the conditional probability: the probability of event B occurring given that A is true. It is also called the posterior probability of B given A.

In this way, the prior knowledge of some events or parameters can help obtain new probabilities, i.e., posterior information.

Bayesian statistics is an approach to data analysis based on Bayes' theorem, where available knowledge and experience about parameters in a statistical model is updated with the information in observed (proxy) data. The background knowledge is expressed as a prior distribution and combined with observational data in the form of a likelihood function to determine the posterior distribution. The posterior can also be used for making predictions

about future events [Schoot et al., 2021]. In them, you use probability to represent all uncertainty within the model, both the uncertainty regarding the output but also the uncertainty regarding the input (parameters) to the model.

$$\underbrace{p(\boldsymbol{\theta}\,|y)}_{\text{Posterior}} \propto \underbrace{p(y|\theta_1)}_{\text{Data Likelihood}} \underbrace{p(\theta_1|\theta_2)}_{\text{Process Likelihood}} \underbrace{p(\theta_2)}_{\text{Prior}}$$

Figure 1: Structure of a Bayesian model.

2 Proposed framework

2.1 General model structure

The general structure of the space-time Bayesian hierarchical model is presented followed by specifics of its application to tropical Pacific SSTs.

Consider the climate field, $Y(\mathbf{s},t)$ at several (m) locations $\mathbf{s}$ and time t.

The joint distribution of the field, $Y(\mathbf{s},t)$ at m locations at each time t is modeled as a realization from a multivariate normal distribution ($\mathcal{MVN}$). The $\mathcal{MVN}$ is characterized by a vector of means, $\boldsymbol{\mu}(\mathbf{s},t)$, that vary over time, and a covariance matrix $\boldsymbol{\Sigma}(\mathbf{s})$. Spatial dependence is captured through spatial and temporal processes on the mean $\boldsymbol{\mu}(\mathbf{s},t)$ and standard deviation $\boldsymbol{\sigma}(\mathbf{s})$ parameters at each location. The latter does not depend on time, strategy used in other models such as the one presented in Tingley and Huybers [2010]. The first data layer of the hierarchical model structure is:

Data layer

$$[Y(\mathbf{s_1},t),\ldots,Y(\mathbf{s_m},t)] \sim \mathcal{MVN}[\boldsymbol{\mu}(\mathbf{s},t),\boldsymbol{\Sigma}(\mathbf{s})] \tag{2}$$

where $\mathcal{MVN}$ stands for "m-dimensional multi variate normal distribution" with dependence matrix $\boldsymbol{\Sigma}(\mathbf{s})$.

The second layer of the hierarchy, the process layer, involves spatial models for the parameters of the normal distribution. The mean vector $\boldsymbol{\mu}(\mathbf{s},t)$ is assumed to vary in time as a function of covariates $\mathbf{Z}(t)$ given by:

Temporal Process layer

$$\boldsymbol{\mu}\left(\mathbf{s},t\right) = \alpha_0\left(\mathbf{s}\right) + \sum_{i=1}^{n_{pc}} \alpha_i\left(\mathbf{s}\right) \cdot Z_i\left(t\right) \tag{3}$$

The standard deviation is allowed to vary only in space.

The regression coefficients $\alpha_i\left(\mathbf{s}\right)$ corresponding to each covariate at each location $\mathbf{s}$ is modeled as a function of vector of spatial covariates –$\mathbf{x}_{\gamma}^{\mathbf{T}}\left(\mathbf{s}\right)$– latitude and longitude at each location. The residual component $\omega_{\gamma}\left(\mathbf{s}\right)$ follows a mean 0, stationary, isotropic Gaussian process (GP) with covariance function $C_{\gamma}(\mathbf{s},\mathbf{s}')$ in where γ is any of the parameters $(\alpha_i,\boldsymbol{\sigma})$. The spatial layers are given as:

Spatial layer for mean

$$\alpha_0\left(\mathbf{s}\right) = \beta_{\alpha 00} + \mathbf{x}_{\alpha 0}^{\mathbf{T}}\left(\mathbf{s}\right)\boldsymbol{\beta}_{\alpha 01}\left(\mathbf{s}\right) + \omega_{\alpha 0}\left(\mathbf{s}\right) \tag{4}$$

$$\alpha_i\left(\mathbf{s}\right) = \beta_{\alpha i0} + \mathbf{x}_{\alpha i}^{\mathbf{T}}\left(\mathbf{s}\right)\boldsymbol{\beta}_{\alpha i1}\left(\mathbf{s}\right) + \omega_{\alpha i}\left(\mathbf{s}\right) \qquad i = 1,\ldots,n_{pc} \tag{5}$$

Spatial layer for standard deviation

$$\boldsymbol{\sigma}\left(\mathbf{s}\right) = \beta_{\sigma 0} + \mathbf{x}_{\sigma}^{\mathbf{T}}\left(\mathbf{s}\right)\boldsymbol{\beta}_{\sigma}\left(\mathbf{s}\right) + \omega_{\sigma}\left(\mathbf{s}\right) \tag{6}$$

where $\beta_{\gamma 0}$ are spatially constant intercept terms, $\mathbf{x}_{\gamma}^{\mathbf{T}}(\mathbf{s_i})$ is a vector of p spatially varying predictors, and $\beta_{\gamma}(\mathbf{s}) = [\beta_{\gamma 1}(\mathbf{s}), \ldots, \beta_{\gamma p}(\mathbf{s})]^T$ is a vector of p spatially varying regression coefficients, γ being one of the parameters presented before.

As mentioned before, $\omega_{\gamma i}(\mathbf{s})$ is the residual at every location and its covariance matrix $C_{\gamma}(\theta_{\gamma})$ as,

$$C_{\gamma}(\theta_{\gamma}) = [C_{\gamma}(s_i, s_j; \theta_{\gamma})]_{i,j=1}^{m} = \delta^2 * \exp -|s_i - s_j|/a_{\gamma}) \qquad i \neq j \tag{7}$$

$$C_{\gamma}(\theta_{\gamma}) = [C_{\gamma}(s_i, s_j; \theta_{\gamma})]_{i,j=1}^{m} = \delta^2 + \tau^2 \qquad i = j \tag{8}$$

The covariance matrix $\mathbf{\Sigma}$ presented in Equation 2 is based on the dependence between sites and is assumed to be a function of distance [Renard, 2011]. The dependence matrix is constructed with a simple exponential model, and has the following shape

$$\mathbf{\Sigma}(\mathbf{s}) = diag(\mathbf{exp}[\boldsymbol{\sigma}(\mathbf{s})]) \cdot \boldsymbol{\rho} \cdot diag(\mathbf{exp}[\boldsymbol{\sigma}(\mathbf{s})]) \tag{9}$$

In which $diag(\boldsymbol{\sigma}(\mathbf{s}))$, is the diagonal matrix $(m \times m)$ with all the standard deviations in every location, expressed in Equation 6.

$$\boldsymbol{\rho}(i, j) = \exp(-|s_i - s_j|/a_0) \tag{10}$$

where a_0 is the range parameter. Note that the values in this dependence matrix $\boldsymbol{\rho}$ are not covariances since they are not scaled by the variance parameters, though the dependence matrix is a valid covariance matrix. By analogy with the variogram, the dependence model is termed the dependogram [Renard, 2011].

Modifications to large domain:

For large domain with many locations the above spatial model gets computationally expensive. To address this, the coefficients are modeled at a few 'knot' locations covering the domain and using Gaussian kernels with radial basis functions. This cuts the computational effort significantly. This idea was proposed by Bracken et al. [2016] and applied to modeling spatial precipitation extremes over the entire western United States. The application domain in this study of tropical Pacific SSTs is several times larger. Incorporating this, the second spatial layer of the hierarchy involves a spatial model for these regression coefficients given as:

$$\beta_{\gamma i}(\mathbf{s}) = \sum_{j=1}^{k} c_{\gamma ij} \eta_{\gamma ij}(\mathbf{s}; a_{\gamma ij}) \qquad i = 1, \ldots, p; \quad \gamma = \alpha_0, \ldots, \alpha_{npc}, \sigma \tag{11}$$

where the γ can represent any of the parameters of the spatial layer, $\beta_{\gamma i}(s)$ is the $i - th$ (spatially varying) regression coefficient, $c_{\gamma ij}$'s are weights for k radial basis functions, the $\eta_{\gamma ij}$'s, which are distributed throughout the domain.

Each regression coefficient is represented as a weighted sum of radial basis functions (Equation 11). The form of these radial basis functions is

$$\eta_{\gamma ij}(\mathbf{s}; a_{\gamma ij}) = \exp(-|\mathbf{s} - \mathbf{s}_j|/a_{\gamma ij}^2) \tag{12}$$

where $a^2_{\gamma ij}$ is a range parameter determining the spatial extent of the basis function. These radial basis functions are Gaussian kernels, which are placed at several points – at the 'knot' locations. The sum of the radial basis functions creates a smoothly varying surface for each regression coefficient.

As mentioned, this technique has been used in Bracken et al. [2016] with good results and computationally more efficient than calculating the spatial parameters at every location of the data field. The model structure described above is graphically represented in Figure 2.

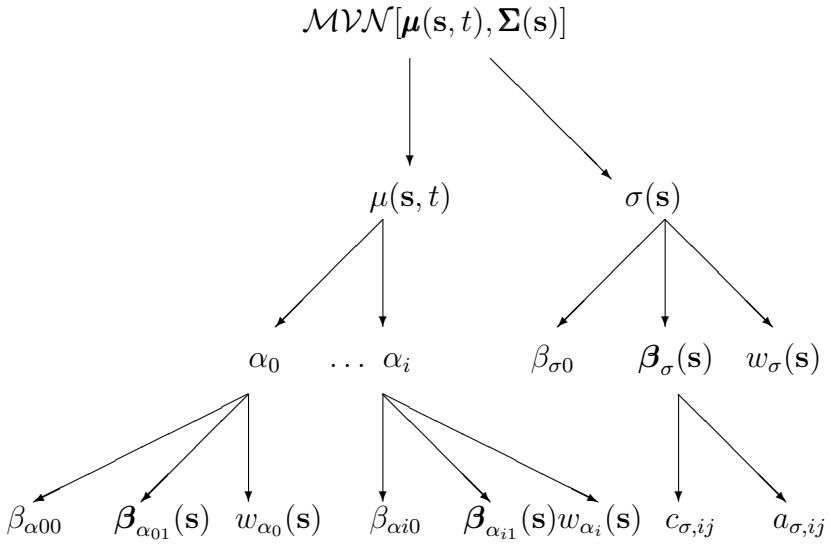

Figure 2: Graphical scheme of the model parameters.

The vector of parameters in the model is

$$\theta = \Big\{ \mu, \sigma, \rho; \alpha_\gamma, \beta_\gamma; c_{\gamma ij}, a_{\gamma ij},\ a_0,\ a_\gamma,\ \delta_{\gamma ij},\ \tau_{\gamma ij} \Big\} \tag{13}$$

The likelihood function of the model can be written as:

$$. \tag{14}$$

Suitable priors are chosen are chosen for the hyper parameters, i.e, the parameters of the lowest layers. A normal distribution has been assigned to the parameters that correspond to coefficients and uniform distributions have been assigned to the parameters that correspond to range parameters. Using Markov Chain Monte-Carlo (MCMC) implemented in STAN the posterior distribution of the parameters is obtained and consequently the predictive posterior distribution of the field at any spatial location and time.

If the field to be modeled requires a distribution other than Normal – such as Gamma for a positive variable or Generalized Extreme Value distribution for block maxima – then the marginal distribution at each location is modeled with a suitable distribution and a Copula is applied to capture the spatial dependence (e.g., Ossandón et al. [2021], Bracken et al. [2016], Bracken et al. [2018]).

3 Application to the Equatorial Pacific SST

The BHM framework proposed in the previous section is demonstrated by its application to modeling and paleo reconstruction of average yearly SST over equatorial pacific. The spatial domain for the application in this study is 10°S to 10°N and 100°E to 75°W (285°E) – shown in Figure 3. The application domain is gridded in a 2°-by-2° grid. The El Niño Southern Oscillation (ENSO) is a main driver of SSTs on interannual and multi-decadal time scales. The climatological pattern of contemporary SSTs is shown in Figure 6, corresponding to the domain in where the ENSO patterns of SSTs happen and which have a major impact on weather and climate around the world [Delage and Power, 2020, UN2, NOA]. Therefore, skillful modeling of SSTs and paleo reconstruction is of immense help in understanding the range of variability and consequently, that of the global climate. Our interest is in reconstructing Holocene (0 to 10k years before present) period SSTs.

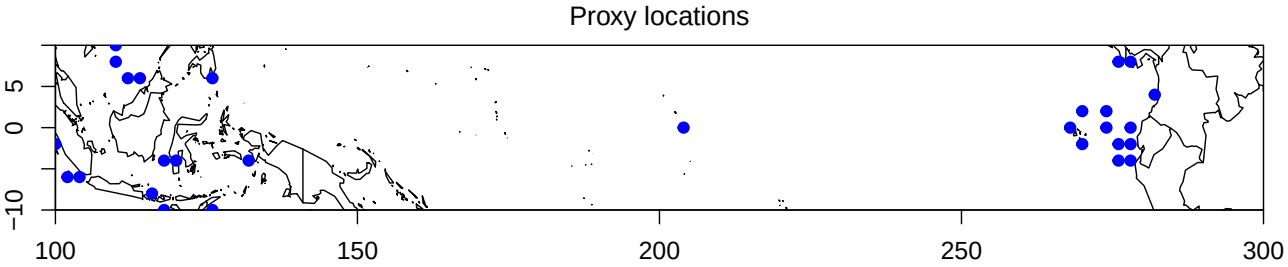

Figure 3: Map of the equatorial pacific, region being studied with the 28 proxy locations.

3.1 Data

The data used for the models is obtained from open sources described below.

3.1.1 Contemporary data

Gridded (2° by 2°) monthly SSTs from 1854 to 2013 are obtained from the NOAA National Climatic Data Center (NCDC) Extended Reconstruction Sea Surface Temperature (ERSST) version 3b data set [Smith et al., 2008] 1981–2010 climatology. Following [Gill et al., 2016] using these, annual (May – April) average SST anomalies are computed. This averaging period is chosen as it captures the ENSO cycle, which typically peaks during the winter months (Nov – Mar).

Excluding the grid points that would fall on land (Indonesia and other Pacific islands) there are a total of $m = 973$ "full field" locations.

3.1.2 Paleo SST Data

The proxy data used in Gill et al. [2016] in western and eastern regions of the domain and, the recent data from central Pacific from Monteagudo et al. [2021] are used in this study.

One common proxy for palaeothermometry exploits Mg/Ca ratios from planktonic foraminifera shells, such as Globigerinoides rubber, extracted from deep-sea sediment cores. At high temperatures, more Mg is incorporated in the shells of foraminifera, and a ratio of Mg/Ca can be used to infer SSTs back through time [e.g., Lea et al. [1999]; Nürnberg et al. [1996]]. Paleo SST records can also be inferred using alkenones produced by some species of coccolithophores (unicellular eukaryotic phytoplankton), such as Emiliana huxleyi. Alkenones are transfat altered under various temperatures: lower temperatures lead to an increase in the degree of unsaturation. The ratio of ditriunsaturated ($C_{37:2}$) to triunsaturated ($C_{37:3}$) alkenones is used to create an Alkenone Unsaturation Index named $U_{37}^{k'}$ ($= C_{37:2}/(C_{37:2} + C_{37:3})$) [e.g., Brassell et al. [1986]; Herbert [2003]], which scales directly to SSTs. Numerous Mg/Ca and $U_{37}^{k'}$ proxy records are available from east and west Pacific [Leduc et al., 2010] and are used in Gill et al. [2016] and other studies. The SSTs from Mg/Ca and alkenone SST records for the west and east equatorial Pacific within the study domain were obtained from the archives of NCDC (https://www.ncei.noaa.gov/products/paleoclimatology/modeling) and Pangaea (http://www.pangaea.de). A recent Mg/Ca proxy record in the crucial region of central Pacific was compiled by Monteagudo et al. [2021] and is included in this study.

Records only with temporal resolutions of at least one value per 1000 years are selected resulting in 28 proxy data at locations indicated in Figure 3. All records have been smoothed using a local polynomial method with a second-order polynomial and a local neighborhood consisting of 70% of the nearest data points [e.g., Loader [1996]] in order to obtain a value for each record representative of every 500 years from now until 10k years ago. Note that smoothed records that stop prior to 2 ka are plotted but were removed from the spatial field reconstructions. All records that included data equal to or younger than 2 ka B.P. were smoothed to 0 ka. Records were then converted to SST anomalies using the 0 ka temperature from each smoothed proxy record.

3.2 Model for the Equatorial Pacific

For the data layer, the annual SST anomalies at each of the m=973 grid points and in each year is assumed to be MVN. The distributions of the SST anomalies during the contemporary period are fairly Normal (see Probability Density Function (PDFs) of SSTs from few locations in Figure S1). The covariates in the temporal process layer (Equation 3) are the leading three Principal Components (PC1, PC2 and PC3) of SSTs obtained from a Principal Component Analysis (PCA) of the contemporary SSTs at the 28 proxy locations – i.e., 'limited field'. The PCA has been performed on both the full field and limited field. From the Eigen spectrum, it can be seen that the leading three leading principal components (PC1, PC2 and PC3) of the limited field explained most of the data variance (80%, 10%, and 5% respectively) and the full field (68%, 10% and 7% respectively). Which sums up to a 95% of the total variance explained for the limited field and 85% for the full field, as can be seen in Figure 4. Considering more than 3 PCs would lead to overfitting in the parameters.

The leading PCs of the limited field have been correlated with the SSTs at all the grid points in the equatorial Pacific, shown in Figure 5. The first PC show strong correlation in the equatorial and eastern Pacific regions – reminiscent of the classic ENSO pattern. The second PC is highly correlated in the western Pacific region, while the third, mostly in central Pacific. This strongly indicates that the information from limited field can be good

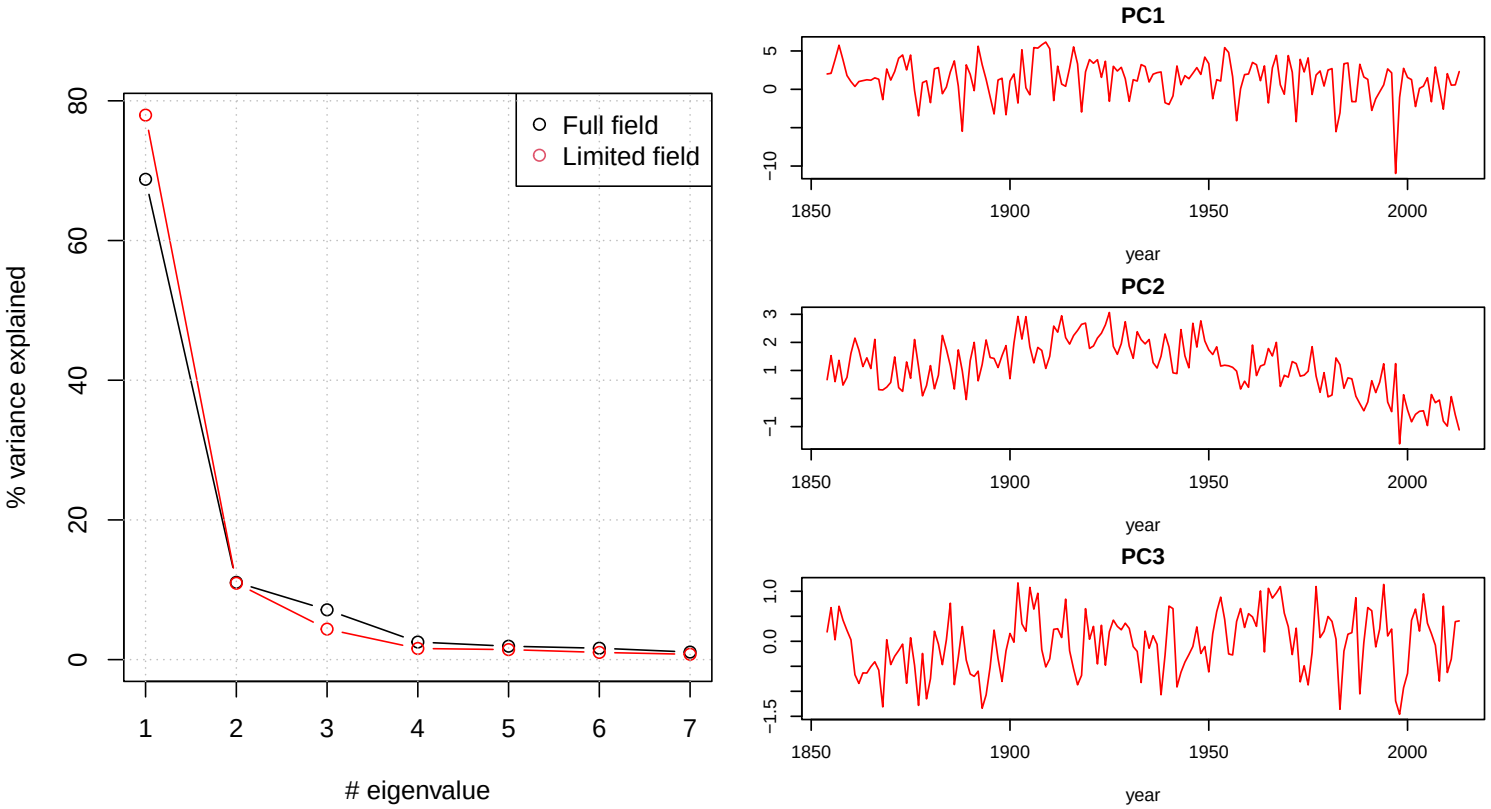

Figure 4: Left: Eigenvalue spectra for (a) the full SST field (1854–2013; black), and (b) limited SST field (1854–2013; red). Right: Time series of the first 3 PCs of the limited field.

predictors for modeling the full field. Thus, in the temporal layer, Equation 3, the covariates **Z** are the three PCs of the limited field.

3.3 Spatial regression knot locations

To ease the computational burden of modeling jointly the SSTs over the entire Pacific, we selected 18 knot locations (Figure 6). These are the locations where radial basis functions or Gaussian kernels are placed (Equation 12). The sum of the radial basis functions creates a smoothly varying surface for each regression coefficient.

The knots are equally placed according the morphology of the ocean. For every information cluster 6 knots have been placed in an organized manner, and respecting the coastal geography, mostly taken into account in the western Pacific. For each parameter of the distribution, these 18 knots have been used in Equation 11. For simplicity, the same knot locations were used for each parameter though each of them could use different configurations of knots.

This technique has been used in Bracken et al. [2016], and in other papers with good results and much more computational efficiency than calculating the spatial parameters at every location of the data field.

With this, the 973 points of the full field can be represented with 18 knot points which

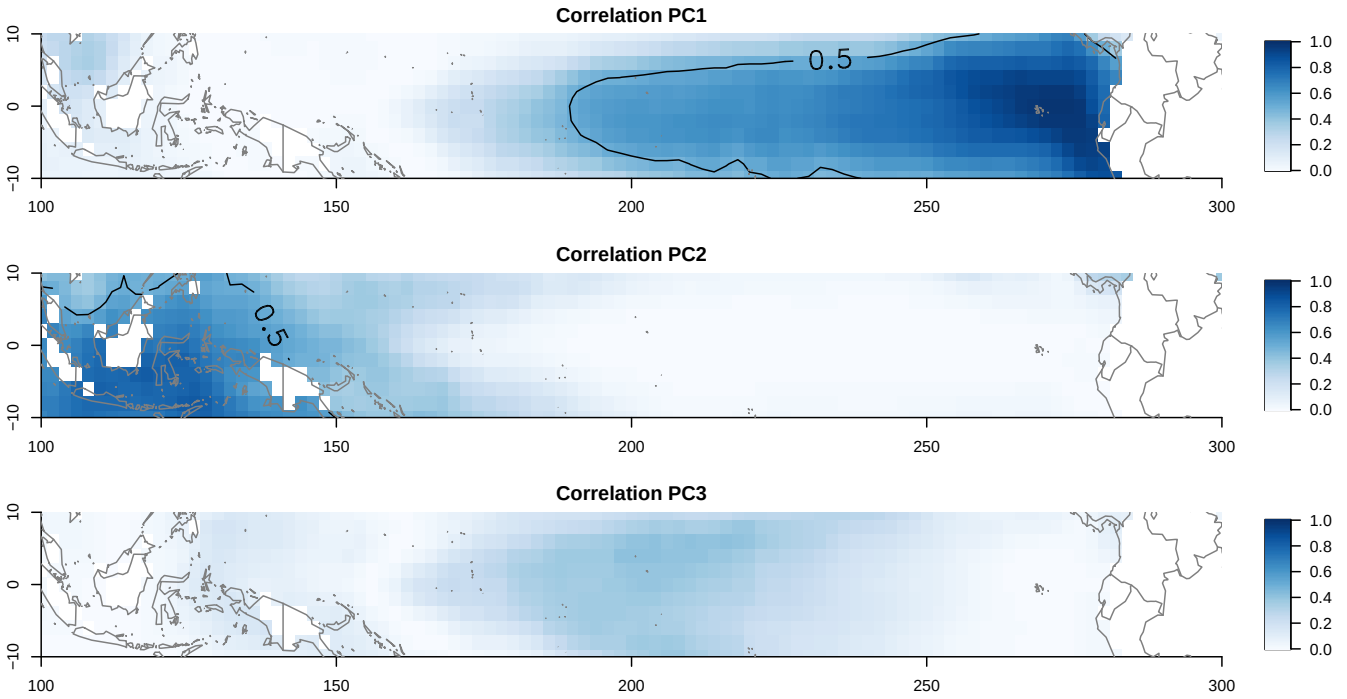

Figure 5: Correlation map of leading three PCs of the limited field with the SSTs of the full field.

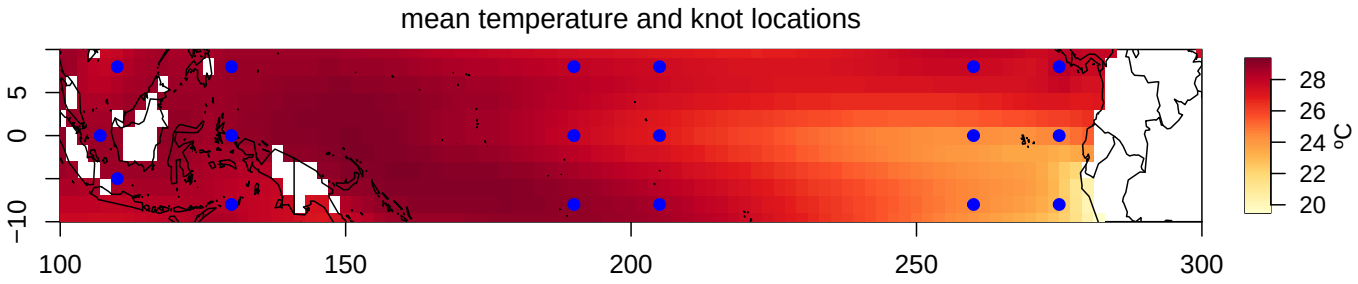

Figure 6: Mean temperature and locations of the spatially varying regression knots.

is 50 times fewer points to fit, thus, reducing the computational time significantly.

3.4 Implementation and model fitting

Candidate BHMCs were fit using the program STAN (Stan Development Team, 2014), its R extension (Stan Development Team, 2020), and the priors assigned in the previous section are suited probability distributions according to the dimensions and scope of the model. STAN uses a Markov chain Monte Carlo (MCMC) simulation method that considers a No-U-Turn Sampler (NUTS; Hoffman and Gelman [2011]) to simulate the posterior probability distribution of the MVN regression coefficients and the dependence matrix. Three parallel chains with different initial values were run, and each chain has a length of 20,000 simulations (iterations) with a burn-in size of 10,000 to ensure convergence. To reduce the sample

dependence (autocorrelation), a thinning factor of 5 has been chosen. The convergence of the posterior distribution of each regression coefficient was checked using the scale reduction factor, , proposed by Gelman and Rubin [1992] – values lower than 1.3 for all the regression coefficients suggest model convergence. Consequently, the posterior distributions of the $\mathcal{MVN}$ regression coefficients, the dependence matrix, and the predictive posterior distributions of annual SST consisted of 6,000 ensembles.

3.5 Model validation and verification metrics

To test the out-of-sample predictability of the model, a validation by dropping 116 years from the record (1854–1969) was performed, and the BHM was fitted using the remaining (training) years (1970-2013). Then, the fitted model is applied to provide estimates for the dropped years.

To assess the at-site performance of the proposed BHM in modeling the observed SSTs, we considered two metrics are computed at each location and displayed as spatial maps:

- The first is the squared correlation coefficient (R^2). This is calculated with the following expression:

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} \tag{15}$$

  In which y is the contemporary data, $\hat{y}$ is the median of the reconstructed data for the full period of each data set (1854-2014) and $\bar{y}$ is the mean of the contemporary data at that specific grid point. For a perfect fit, one would expect $R^2 = 1$ and a baseline model, which always predicts $\bar{y}$, will have $R^2 = 0$.

- The second is the "resolved variance" statistic β, given by the following expression:

$$\beta = 1 - \frac{\sum (y - \hat{y})^2}{\sum y^2} \tag{16}$$

  where y is the contemporary data and is the median of the reconstructed data for the full period of each data set (1854-2014). We compute this statistic at each grid point over the reconstruction domain. For a perfect fit, one would expect $\beta = 1$, and for two random series, one would expect $\beta = -1$.

To assess the temporal performance of the BHM some of the Niño indices are calculated and compared with the available data, also the spatial maps of estimates of SSTs from the model for selected El Niño and La Niña years.

3.6 Paleoreconstruction

As mentioned, proxy SST data at $z(= 28)$ locations over the equatorial pacific are available. These data span from 10,000 years ago until present. Although there is some missing data from the proxies, those are smoothed with data of the same time at other proxy locations.

For any desired year in the paleo reconstruction, say 4000 ky, the PCs are obtained using the proxy SST data at the proxy locations and the Eigen vector from the PCA of the contemporary SST data at the proxy locations. The leading three PC values are used as covariates in the BHM to obtain posterior distribution of SSTs at all the grid points.

4 Results

As mentioned in the previous section, two metrics were computed from the model fitting and validation.

Figure 7 shows the spatial map of R^2 and β of the fitting. The R^2 for the fitting is greater than 0.9 at the locations close to a proxy location, as to be expected as the PCs from these are used as covariates in the model. The values are generally above 0.7 in most of the basin expected around the 160^o longitude coordinates. This is mostly due to the lack of proxy data in this region. The performance from our model is better than that of Gill et al. [2016] largely due to the recent proxy data we have from central Pacific that they did not have. In addition, in Gill et al. [2016] they perform the PCA on the contemporary SST data on the full grid ($\sim$ 973 locations) as compared to much less time points ($\sim$ 160), this introduces noise in the Eigen vector estimates. Where in our model here, PCA is performed on SSTs from just the proxy locations ($\sim$ 28) leading to a more stable Eigen vectors.

The β parameter also shows very similar results to the square correlation, with the same spatial strengths and flaws.

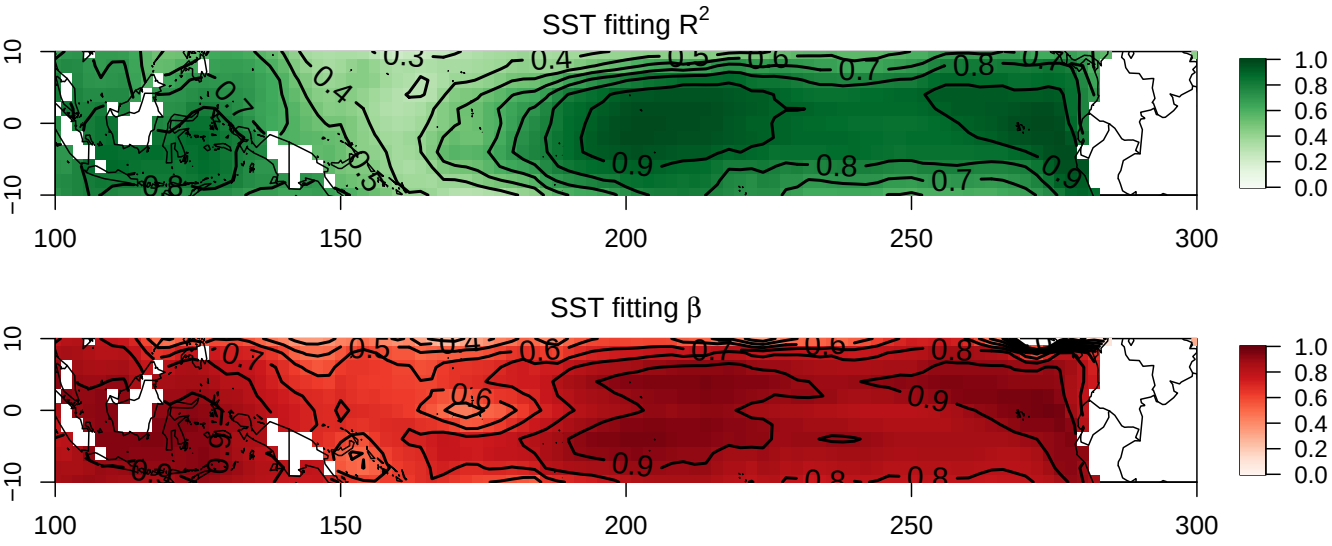

Figure 7: Model fitting and calibration statistics

Model validation was performed by training the model on 1970–2013 data and using that model to validate SSTs (1854–1969). The β and R^2 statistics are computed to quantify model skill and shown in Figure 8.

The spatial distribution of the validation statistics in both coasts is similar to their counterparts from fitting in Figure 7. The skill values are generally higher in those areas (above 0.7), across the basin except for the region around the 160^o latitude. The validation skills are about the same than that of Gill et al. [2016].

To show the performance of the model in modeling selected El Niño (1998) and La Niña years (1955, 1974 and 1989), the posterior median SSTs for these selected years are shown in Figures 9, 10, 11 and 12 along with the historical SST patterns.

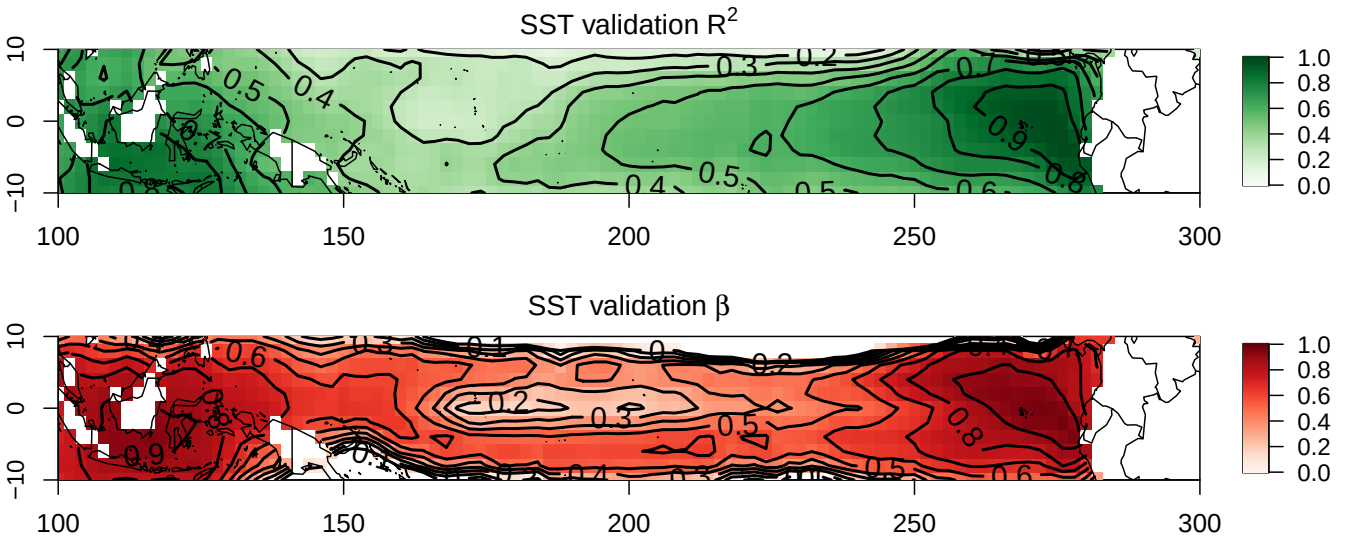

Figure 8: Model validation statistics for the 1854-1969 period of time

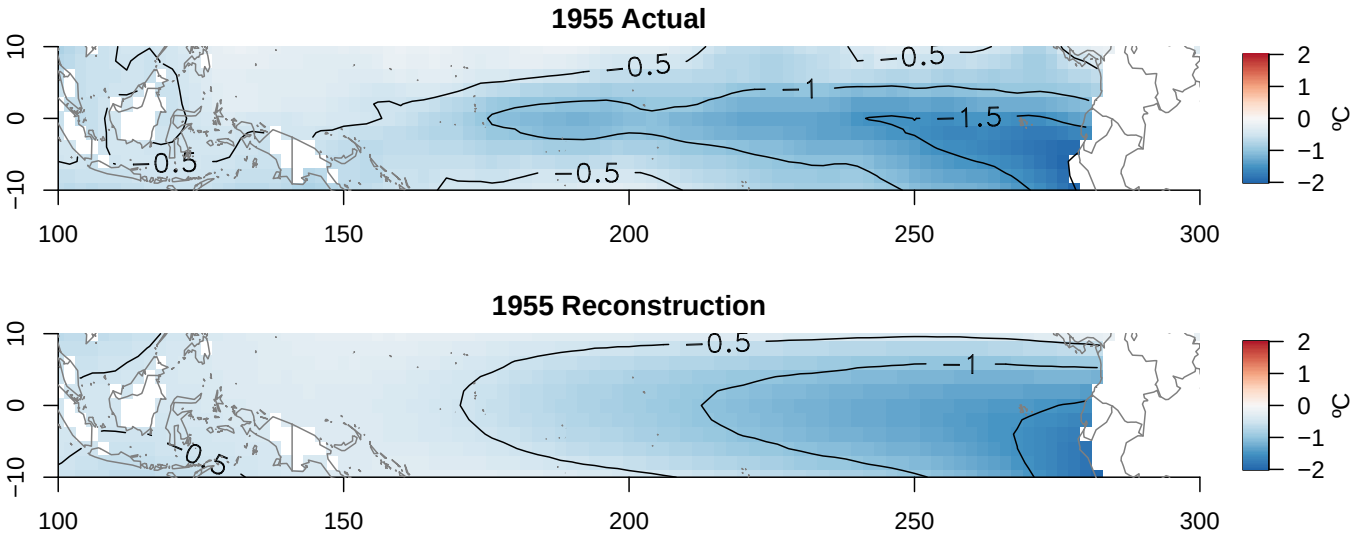

Figure 9: La Niña actual data and median reconstruction of the model in 1955.

The 1955 reconstruction (Figure 9) is very good at capturing the cooling of -1.5^{o}C. However, it falls a little short in reconstructing the full La Niña event up to 240^{o}E, and it stops at 270^{o}E. Similarly in the 1974 (Figure 10) and 1989 (Figure 11) reconstructions the cooling magnitudes are precise, with the spatial extension smaller than the observed.

The reconstructed magnitudes and spatial patterns of anomalous SSTs match well for the strongest El Niño year on record of 1998 (Figure 12) and is better in the spatial pattern compared to the La Niña anomalies presented before. Consistent with the observed warming, the reconstructed anomalous warming falls a little short in extending towards the west; the greatest magnitude of warming, which is off the coast of Peru in the far eastern Pacific,

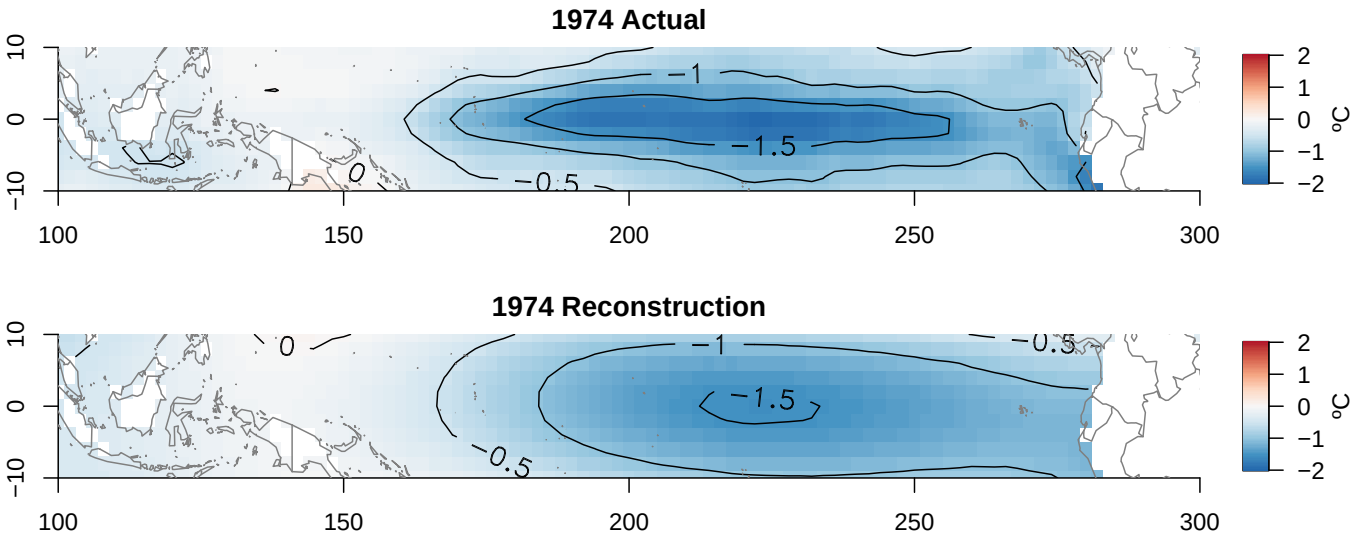

Figure 10: La Niña actual data and median reconstruction of the model in 1974.

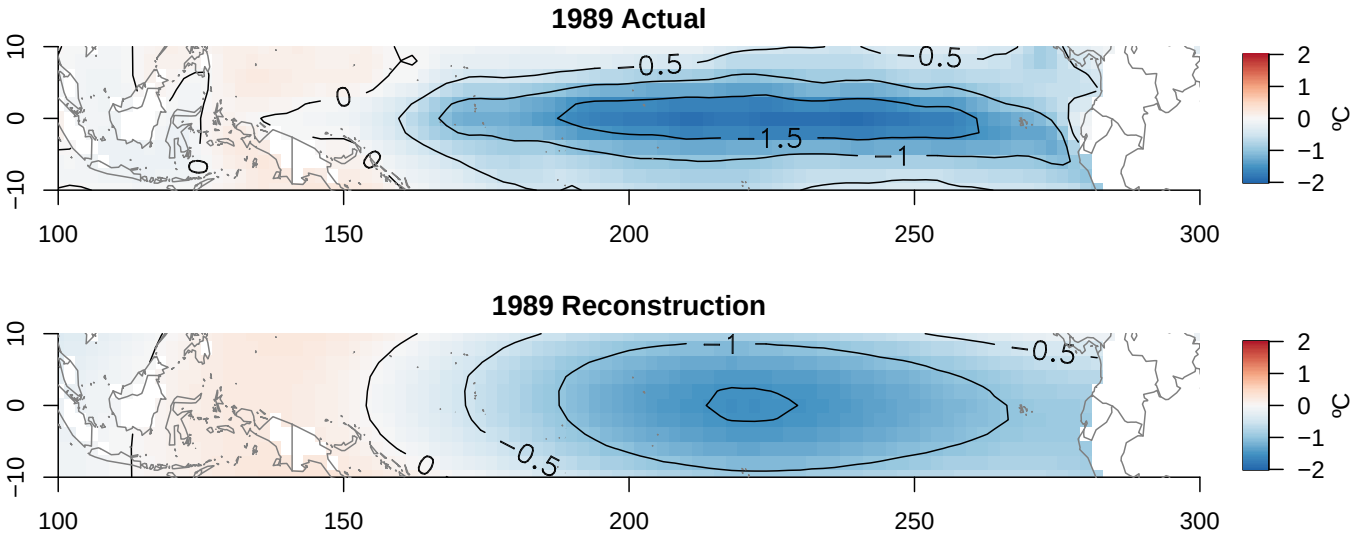

Figure 11: La Niña actual data and median reconstruction of the model in 1989.

reaches approximately $+3.5^{o}$C; and the distribution of temperatures within the reconstructed "warm tongue" closely resemble those of the observed SSTs.

It is likely that the failure of the limited-field model to capture La Niña SST anomalies across the entire Pacific is due to the restricted locations of the proxy records. The cold La Niña anomalies are centered in the east-central Pacific, where there is just one proxy record. The maximum warm anomaly of the 1998 El Niño anomaly, however, is located in the far eastern Pacific, which the palaeoceanographic data sample well. Reconstructed SST for the El Niño year is arguably better than those for the La Niña years as well.

There are several indices used to monitor the ENSO events. These Niño indices are based

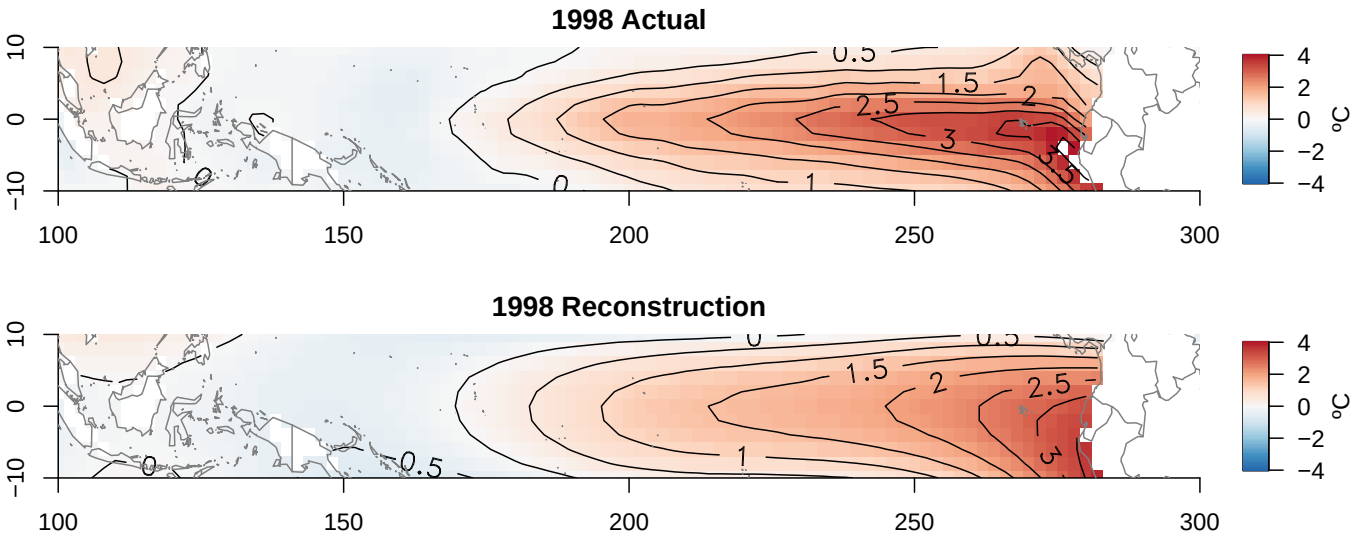

Figure 12: El Niño actual data and median reconstruction of the model in 1998.

on the average SST anomalies over various regions. The Niño 4.0 encloses the central-western tropical Pacific, the Niño 3.4 encloses the central tropical Pacific, and the Niño 1.2 encloses the eastern tropical pacific. The difference between the Niño 1.2 and Niño 4.0 is known as Trans Niño index, TNI. The TNI thus measures the gradient in SST anomalies between the central and eastern equatorial Pacific, capturing El Niño and La Niña events.

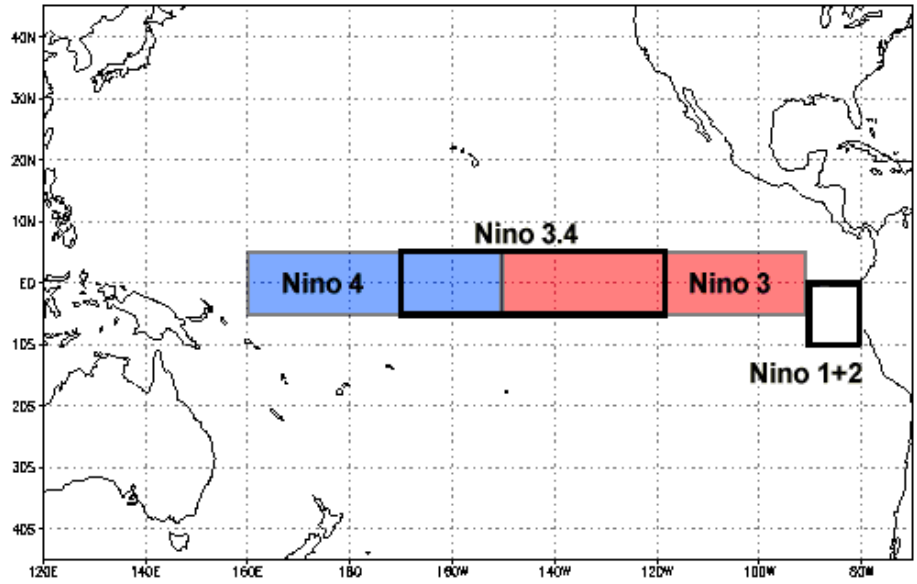

Figure 13: Map of the Niño regions. Source: NOAA

The larger the TNI, the stronger the Niño event is. The distribution of the indices along with historical values are presented in Figure 14.

Throughout the contemporary data, the model Niño indices 3.4 and 4.0 capture the historic values and specifically the ENSO events (i.e., higher values for El Niño and lower for La Nina) very well. The performance of the Niño 3.4, especially for La Niña events (smaller value of the index) is better than the Niño index 4.0. Overall, the indices from the BHM and historic correspond very well along with the attendant uncertainty.

The BHM has been employed to reconstruct SSTs during Holocene. To this end, the SST fields for 10, 8, 6, 4, and 2 ka have been reconstructed, shown in (Figure 15). The SST reconstructions show the mid-eastern Pacific to have been a maximum of about 2^{o}C cooler than today at 10 ka, along with a cold tongue anomaly, cooler by 1^{o}C, extending to about 140^{o}E and 300^{o}E. The western Pacific, however, it is shown to be a little warmer, by 0.5^{o}C, for the same time period, resulting in a maximum zonal difference anomaly of 2 ± 0.30^{o}C. From 10 to 6 ka, the mid-eastern Pacific warmed by about 1.5^{o}C and the western Pacific warmed by about 0.2^{o}C. By 2 ka, most of the domain was within +0.5 to 1 ^{o}C of modern, with a zonal difference just distinguishable from today in the eastern Pacific and far western Pacific. The cooling in the central and eastern Pacific is stronger than what was obtained in Gill et al. [2016]. This is consistent with other studies. Based on compiled coral and mollusk archives much of the Holocene was in a reduced ENSO state [Cobb et al. [2013]; Emile-Geay et al. [2015]], which is consistent with our reconstructions, in that every all the reconstructed SSTs are colder than modern and with larger zonal SST differences.

The same proxy indices as in Figure 14 have been calculated for the Holocene period, and are shown in Figure 16. The trend of the El Niño indices is decreasing and showing two inflection points between 4 ka and 6 ka, resembling what is shown in Figure 15. Other studies [e.g., Gill et al. [2016], Gagan et al. [1998]] show a cooling of the Eastern pacific and a light warming of Western pacific in the mid and late Holocene, captured in these indices.

In Figure 17, the proxy locations used for the limited field are labeled in the different regions.

As a complimentary validation, the reconstructed SSTs at the proxy locations along with proxy derived SSTs are plotted in Figures 26 and 27 in the Appendix. It can be seen that the posterior median of the reconstructions along with the credible intervals capture the proxy SSTs quite well. In fact, the reconstructions also capture the temporal variability well. The locations where there are mismatch are location #12 and #26, which are in both eastern and western regions of the Pacific.

The poor fit of #12 is not surprising, as it shows an overall trend opposite to the rest of the records in the eastern Pacific. Something similar happens with proxy #26 and the trends in the western Pacific. This additional validation also adds confidence in the BHM framework.

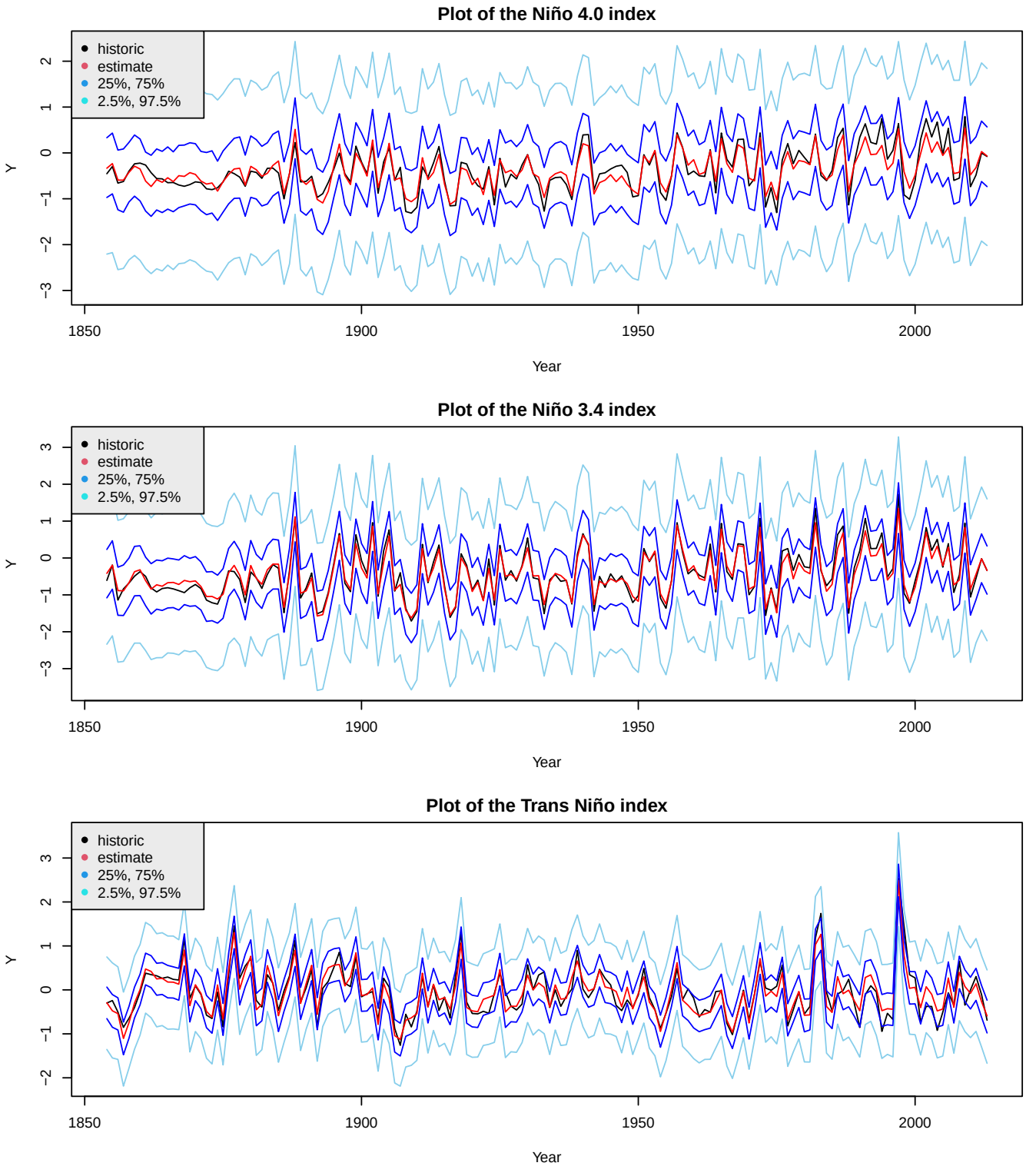

Figure 14: Plot of the reconstructed El Niño indices 4.0 and 3.4 and the Trans Niño Index.

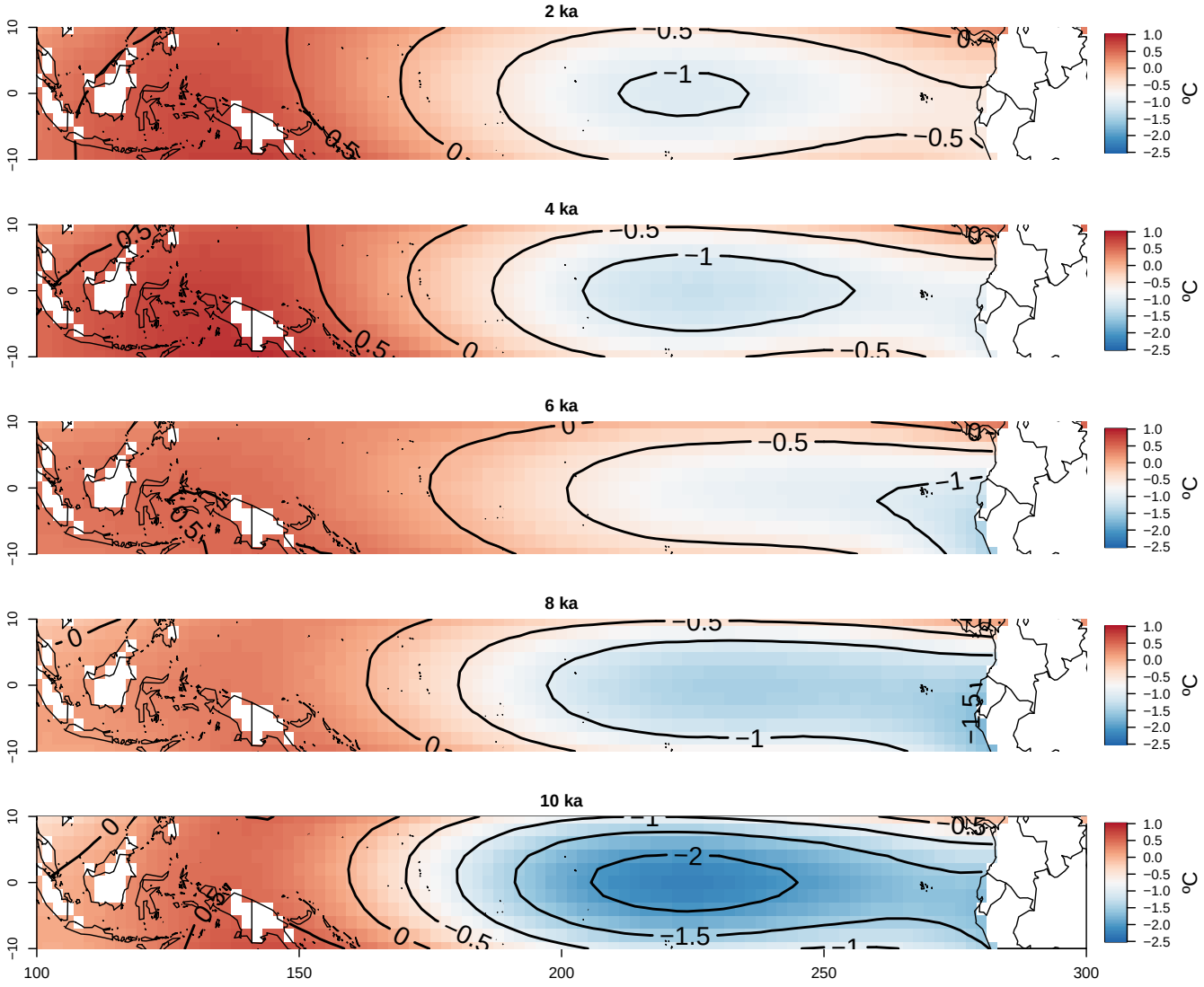

Figure 15: Plot of the median of the reconstructed SSTs during several Holocene years.

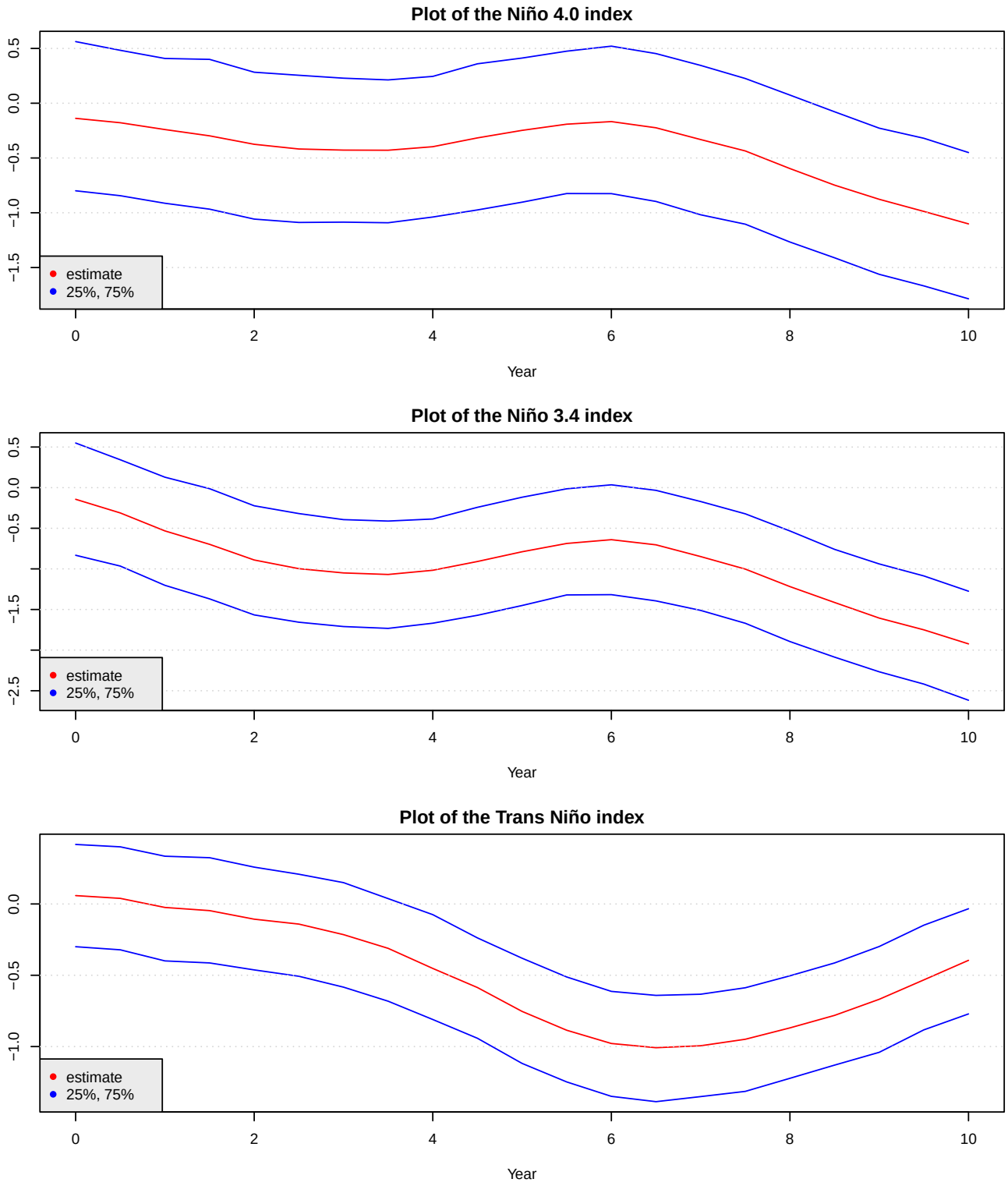

Figure 16: Plot of the reconstructed El Niño indices 4.0 and 3.4 and the TNI for the Holocene period.

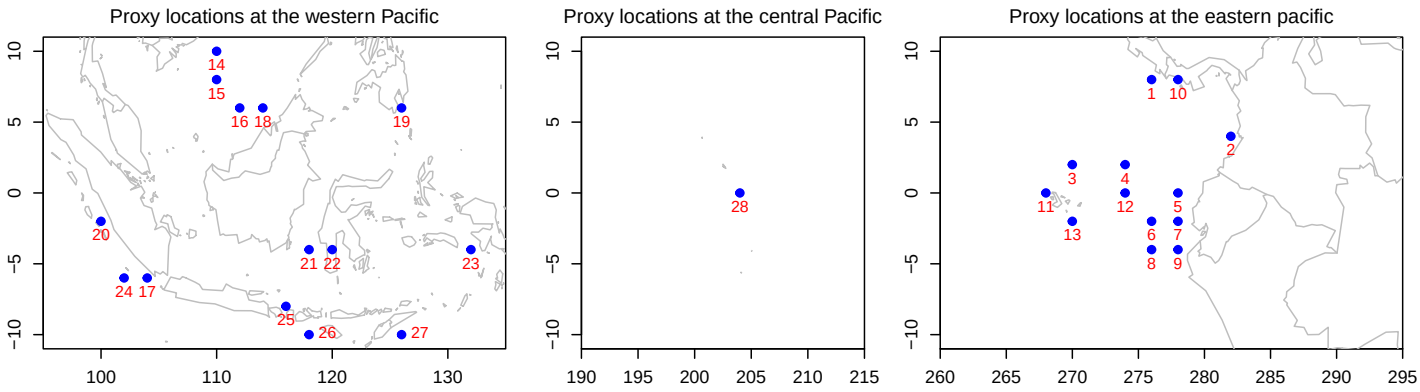

Figure 17: Proxy identification numbers.

5 Semi-Bayesian model

Having seen the Bayesian methodology and the results of the model, it is natural asking if there are simpler methods that could have the similar results and statistics.

The method presented in Gill et al. [2016] shows very little uncertainty (only ad-hoc uncertainty) when trying to reconstruct SST of previous periods of time. Following the same idea, of working with both PCs of the full field and the limited field, introduces us into a semi-Bayesian scheme in which we can at the same time reconstruct the statistics of the past periods of time.

In this case, instead of using the SST, the data are the principal components of the SST over the equatorial pacific over 160 years of data (1854–2014) at each of (G =) 973 grid points spanning the equatorial Pacific from 10^oS to 10^oN and 100^oE to 75^oW (285^oE). They will be referred as the principal components of the full field.

The predictor covariates will still be the principal components of the limited field.

5.1 Formulation of the semi-Bayesian model:

Consider the sea surface temperatures at an ocean grid over the years $Y(\mathbf{s}, t)$. This grid domain is specified and temporally bounded by certain predictors. The semi-Bayesian model models the PCs of the full field with the PCs of the limited field in the following way:

The distribution of the first $n_{pc} = 3$ full field PCs is modeled with normal distributions as follows:

$$PC_{i,ff} \sim \mathcal{N}[\mu_i, \sigma_i] \qquad i = 1, \ldots, n_{pc} \tag{17}$$

where PC_{if} is the $i-th$ principal component of the full field, i.e., these are supposed to have a normal distribution. The mean and the standard deviations are modeled with the principal components of the limited field. The second layer of the hierarchy, also known as the process layer, involves the limited field PCs for the normal distribution parameters,

$$\mu_i = \mu_{i0} + \mu_{i1} \cdot PC_{1,lf} + \mu_{i2} \cdot PC_{2,lf} \tag{18}$$

$$\sigma_i = \sigma_{i0} \tag{19}$$

where the parameters are modeled with the limited field principal components obtained from the SST at the proxy locations for a span of over 160 years of modern data, for every principal component of the full field ($PC_{i,ff}$).

Variance is stationary, in Equation 19, as it stays pretty much constant over time. This is used in other studies, such as Tingley and Huybers [2010].

The prior distributions of the paramaters is assumed to be,

$$\gamma_{ij} \sim \mathcal{N}[0, 1] \qquad \gamma = \mu, \sigma; \;\; i = 1, \ldots, n_{pc}; \;\; j = 0, \ldots, 2 \tag{20}$$

5.2 Results of the semi-Bayesian model

Although this model (sBHM) is much simpler and computationally less demanding, the results can be measured in the exact same way and even be compared with the ones of the BH model.

The R^2 and β statistics are also computed for the fitting and validation of the model, with the same exact data sets as used before.

The β metric is clearly affected by the PCs, as the semi Bayesian model is only able to be fit with the information of the PCs of both limited and full field. The R^2 is comparably good, slightly worse than the one presented in Figure 7.

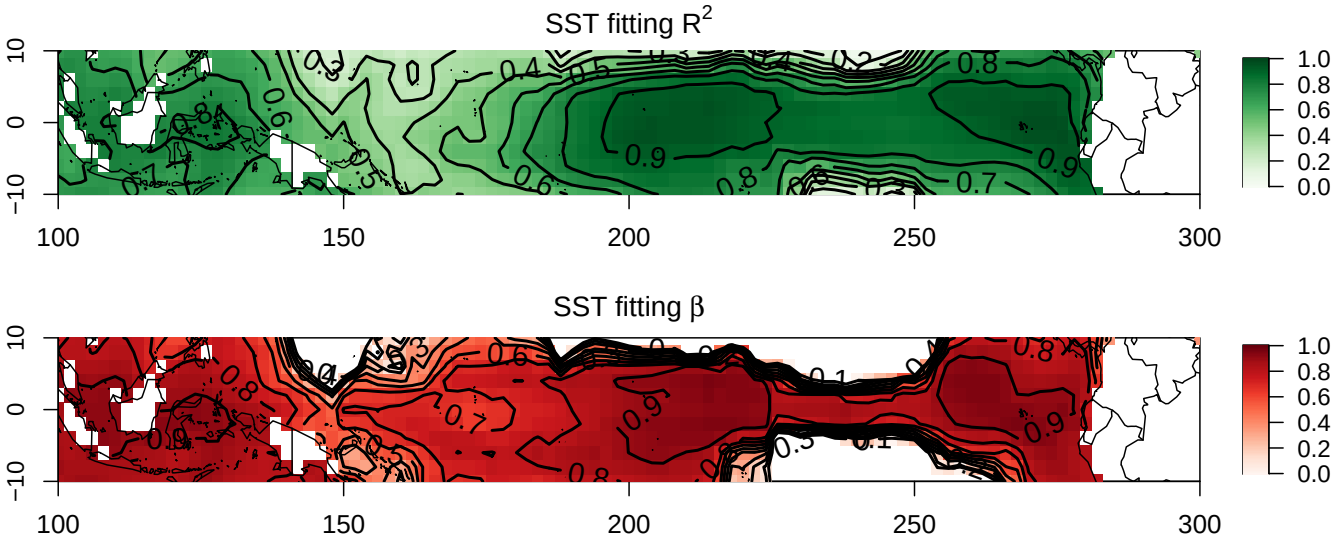

Figure 18: Model fitting and calibration statistics

Model validation was also performed by training the model on 1970–2013 data and using that model to validate SSTs (1854–1969). The β and R^2 statistics are computed to quantify model skill and shown in Figure 19.

Here, again, the β metric is less than zero at the locations distant from the proxy locations, thus, far from the predictors information. The R^2 correlation for the validation model, it is very similar to the one in Figure 8. This difference between R^2 and β is mainly because the standard deviation at each point is bigger than the sum of the squares of the temperature at those points.

The same years of El Niño and la Niña events as before have been selected to reconstruct the temperatures.

The 1955 reconstruction (Figure 20) is very good at capturing the cooling of -1.5^{o}C. However, it falls a too long in reconstructing the full La Niña event in the 190^{o}E, mainly due to the big influence of the third PC of the limited field in that area. Similarly in the 1974 (Figure 21) and 1989 (Figure 22) reconstructions the cooling magnitudes are precise, with the spatial extension a little smaller than the observed.

The reconstructed magnitudes and spatial patterns of anomalous SSTs match well for the strongest El Niño year on record of 1998 (Figure 23) and is slightly better in the spatial

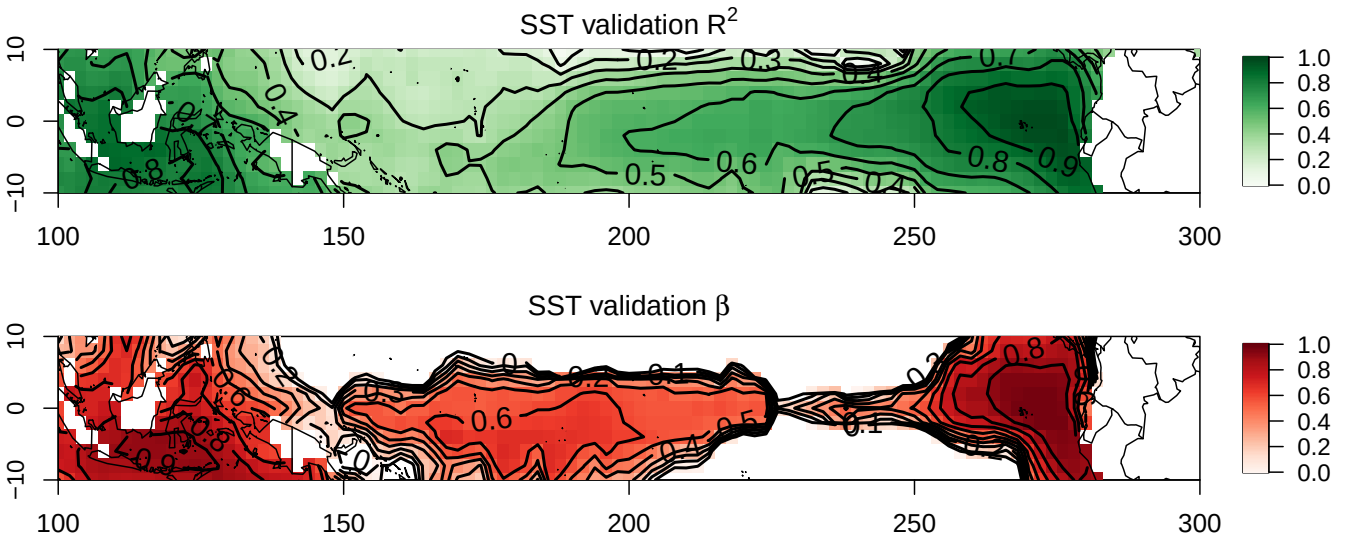

Figure 19: Model validation statistics for the 1854-1969 period of time

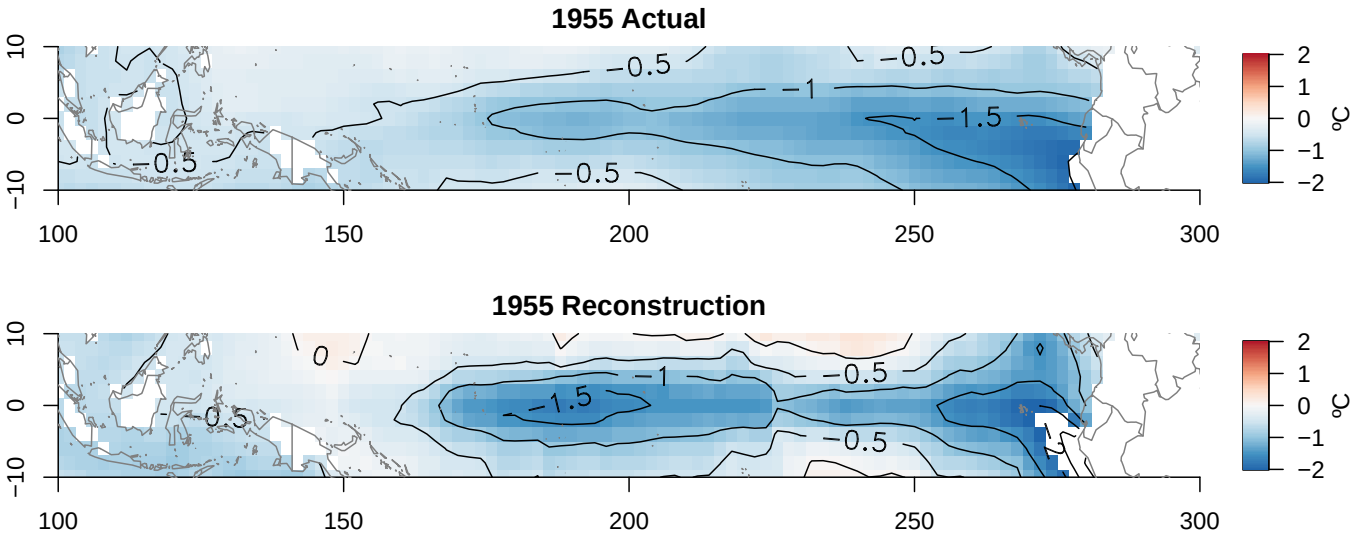

Figure 20: La Niña actual data and median reconstruction of the semiBH model in 1955.

pattern compared to the La Niña anomalies presented before, although it still presents a big influence from the third PC of the limited field. Consistent with the observed warming, the reconstructed anomalous warming falls a little short in extending towards the west; the greatest magnitude of warming, which is off the coast of Peru in the far eastern Pacific, reaches over +3.5ºC; and the distribution of temperatures within the reconstructed "warm tongue" closely resemble those of the observed SSTs, in this case, forming a second tongue.

It is likely that the failure of the limited-field model to capture El Niño and La Niña shapes of the SST anomalies across the entire Pacific is due to the restricted locations of the proxy records, and how the principal components have much bigger importance than in the

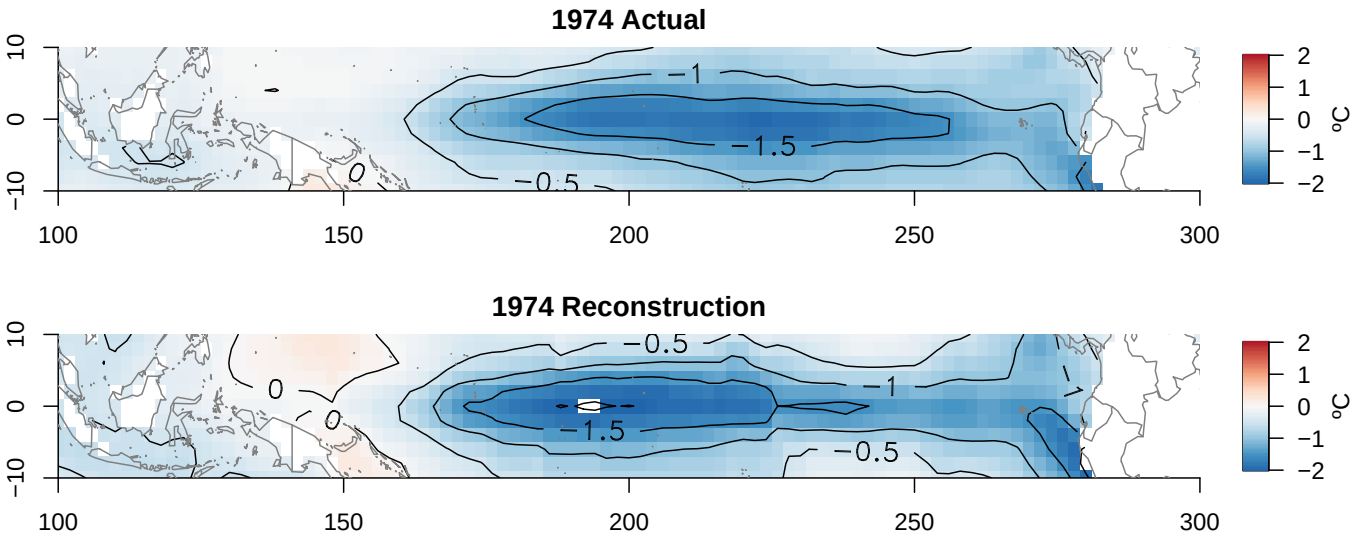

Figure 21: La Niña actual data and median reconstruction of the semiBH model in 1974.

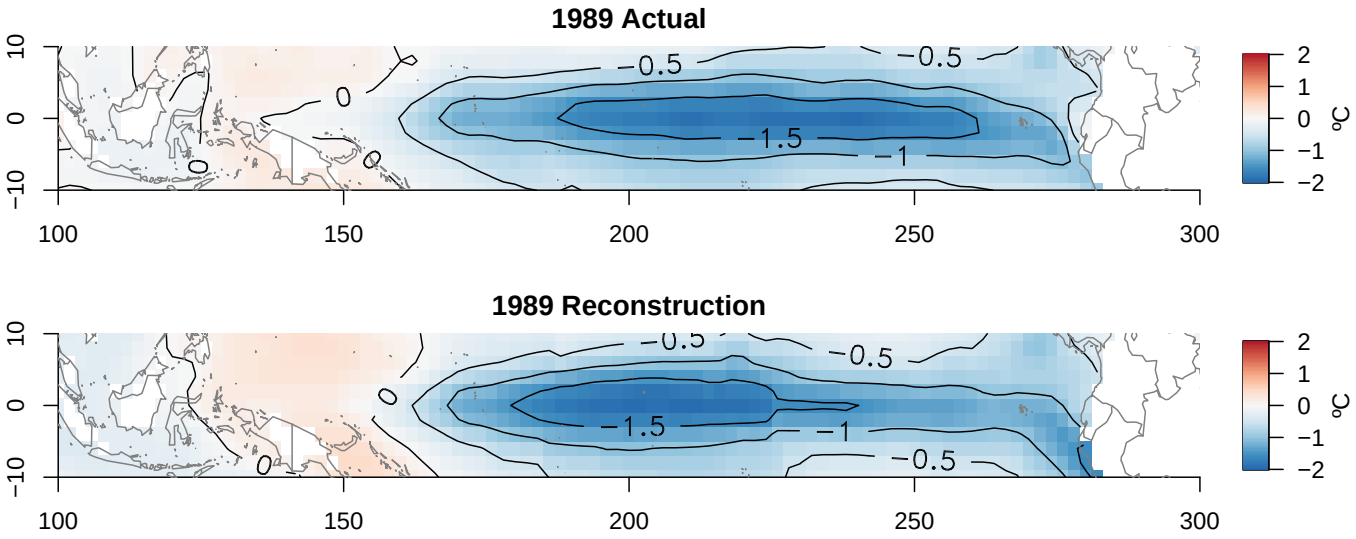

Figure 22: La Niña actual data and median reconstruction of the semiBH model in 1989.

much smoother reality.

Similarly, for the Niño indices, it can be seen how close they are to the historic data, in Figure 24, but at the same time, how the uncertainty has been drastically reduced. At the same time, capturing well both El Niño and La Niña events.

The plots of the median of the reconstruction at 10, 8, 6, 4, 2 ka in Figure 25 is very similar to the one presented before in Figure 15. It presents slightly colder temperatures in the eastern-central Pacific. Reinforcing the results of the Bayesian model, the eastern Pacific cools to -2^{o}C in the early Holocene.

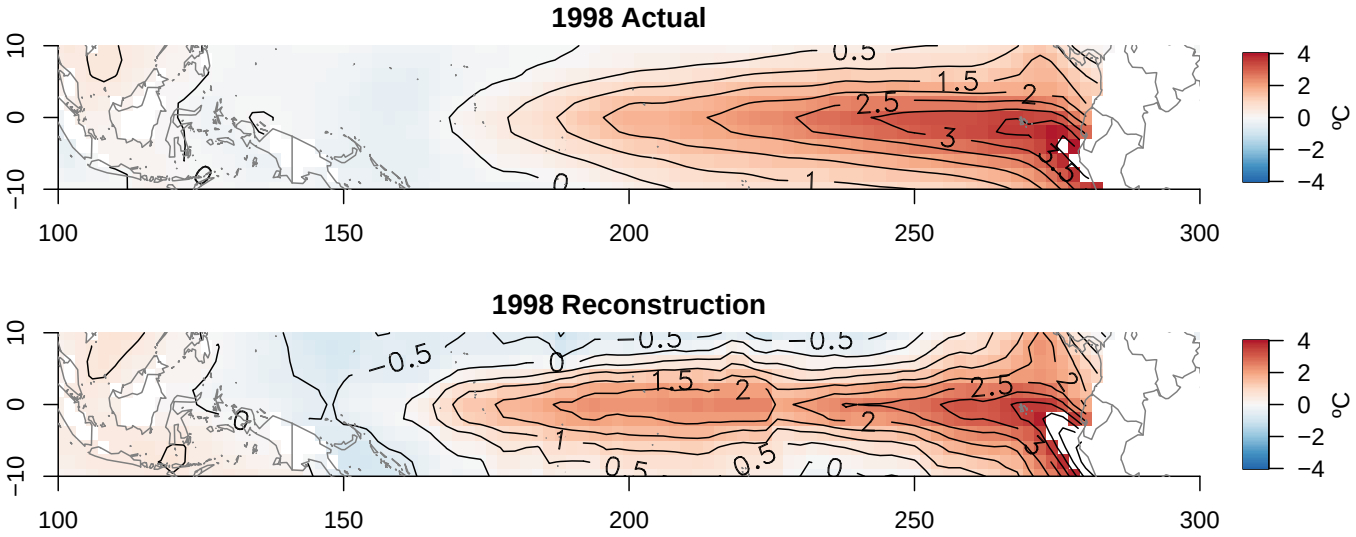

Figure 23: El Niño actual data and median reconstruction of the semiBH model in 1998.

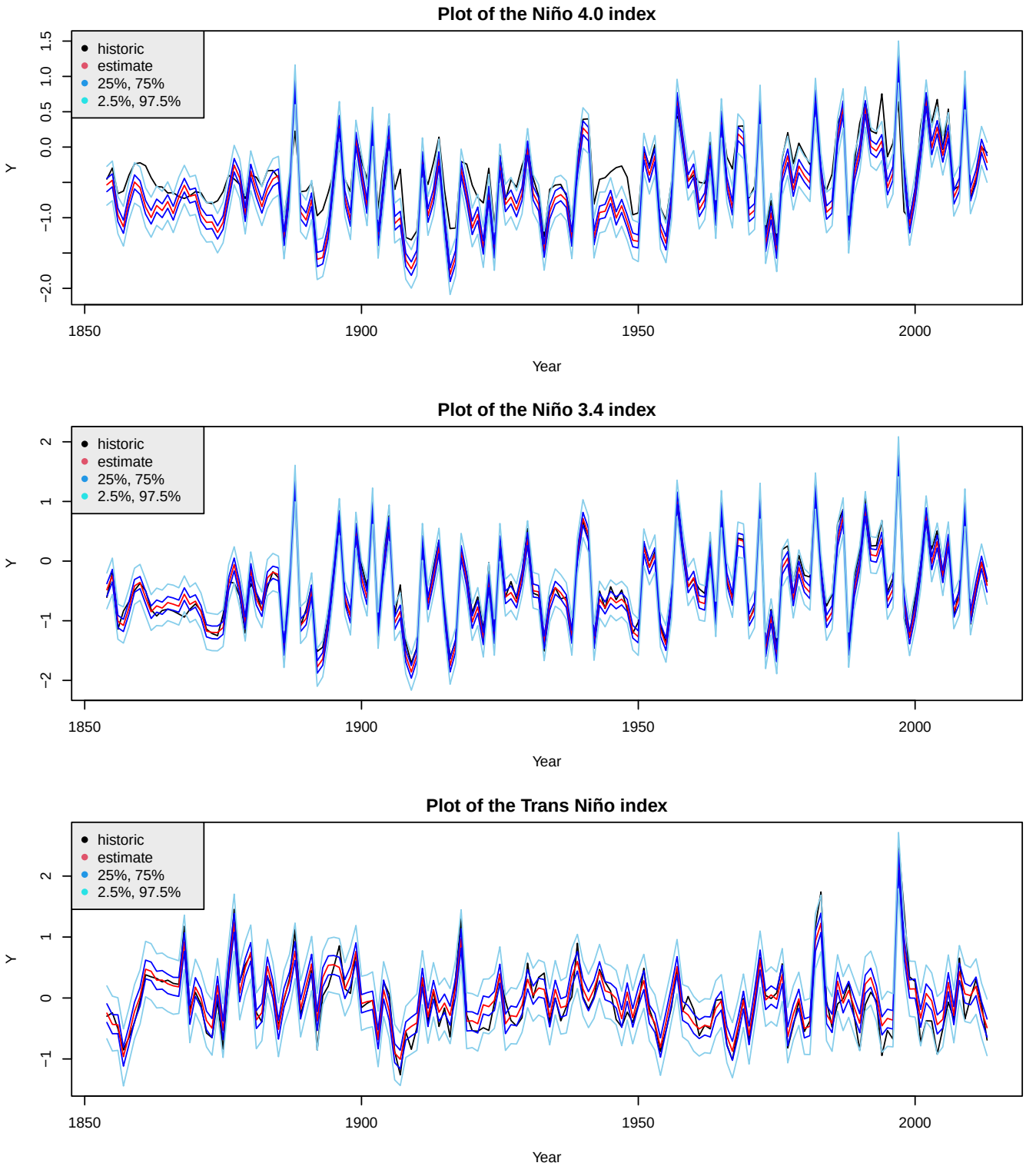

Figure 24: Plot of the reconstructed El Niño indices 4.0 and 3.4 and the Trans Niño Index for the semi Bayesian model.

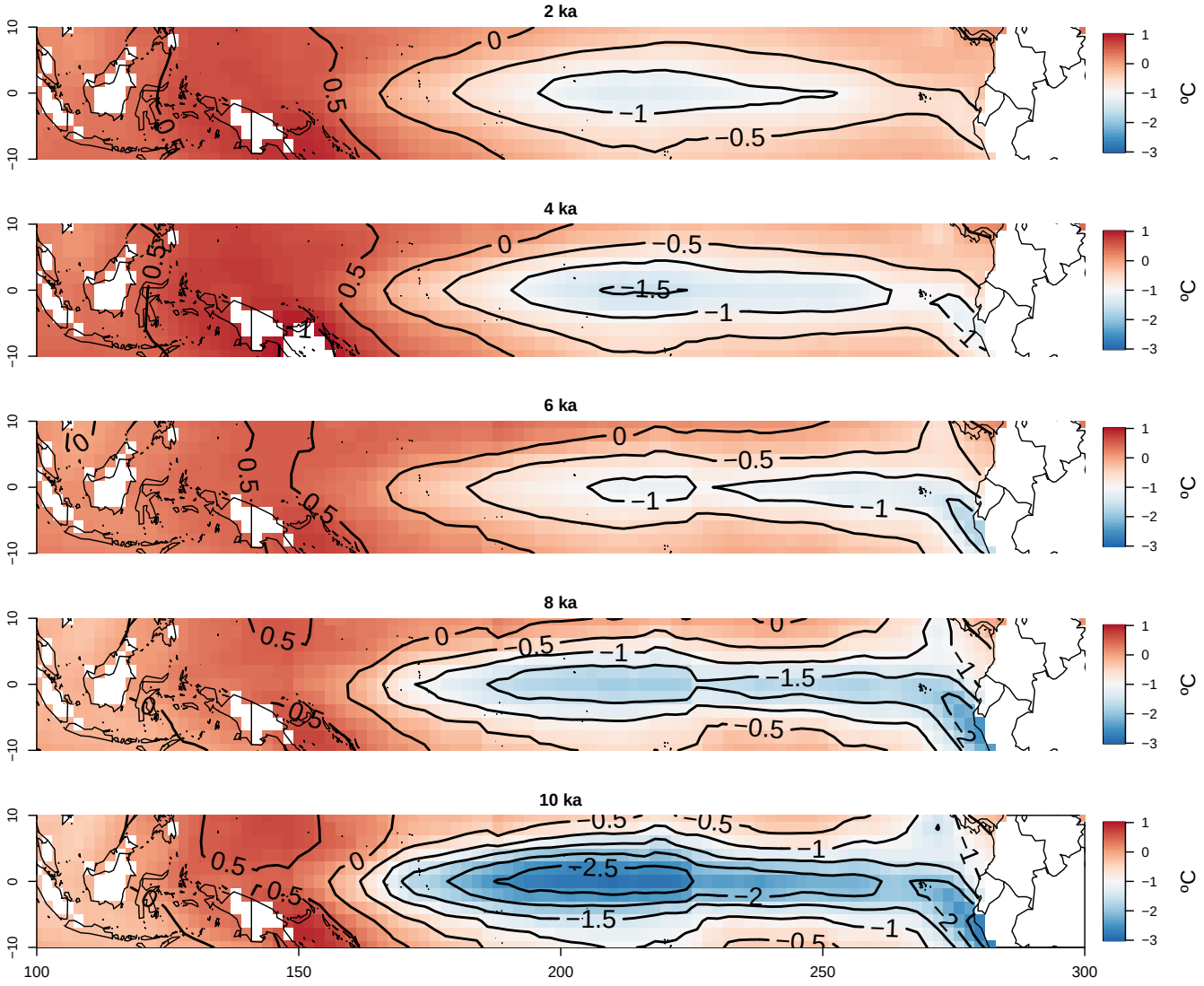

Figure 25: Plot of the median of the reconstructed SSTs during several Holocene years for the semi Bayesian model.

6 Comparison

The semi-Bayesian model presented in the previous section is much more efficient and faster in terms of computational cost. However, the presented results are comparable to the ones corresponding to the Bayesian Hierarchical model.

The main advantages and disadvantages of the BHM are:

Advantages	**Disadvantages**
Able to capture more dependence and variance due to working with the SSTs directly	Bigger computational model
Greater capturing of uncertainties	Slower convergence
The spatial regression adds computational efficiency and flexibility	Restricted to the covariate information

Table 1: Advantages and disadvantages of the BHM.

On the other hand, the main advantages and disadvantages of the semi-BHM are:

Advantages	**Disadvantages**
Simple model easy to define and flexible to adapt to the data sets	Not able to capture extra dependence and covariance, due to working with the PCs of both fields
Faster convergence	Lower ability to capture uncertainties
Results comparable to BHM	Even more restricted to the covariate information

Table 2: Advantages and disadvantages of the semi-BHM.

having seen this, and depending in which future applications these models have to implemented to, one can choose one or the other. It is true that the BHM presents smoother and more robust results, first because the Temperature field is usually a smoother function with greater uncertainty. Also, because it is able to work directly with the objective variable observations and not its principal components.

7 Conclusions

A space-time Bayesian Hierarchical Modeling framework has been developed for climate fields over a large domain. The BHM model has four layers along with a modification to modeling on large spatial domain with computational efficiency. The covariates to capture temporal variability are the leading principal components of SSTs at few locations (~ 28) which also coincide with locations of proxy SST data. The model was applied to annual SSTs over the domain of equatorial Pacific. The posterior median from the model captured the space-time variability of SST over the entire domain very well including spatial patterns of selected ENSO years. The model has been applied to reconstruct SSTs over the domain during Holocene (0 to 10 ka). The results show progressive warming in the central and eastern equatorial Pacific from 10K to present consistent with increased El Niño frequency. There is mid-Holocene cooling during 4 to 6 ka. The greatest eastern equatorial Pacific SST anomaly (cooler by 2^oC) occurred at 10 ka, with gradual warming persisting until 6 ka. By 2 ka, the Pacific remained in a state cooler than present with SST anomalies around -1^oC in the eastern pacific and about $+0.5^oC$ warmer in the western Pacific.

Is has been presented the powerful aspect of the BHM model is its ability to capture the uncertainties robustly, compared to traditional PCA-based techniques such as in Gill et al. [2016] or to the semi Bayesian model presented in this same study. This framework offers the potential to reconstruct other fields such as rainfall, vegetation, etc. Especially rainfall patterns over India can be evaluated with uncertainties - updating and complementing Gill et al. [2017].

The period of time can be also modified, and one can use the BHM to reconstruct any field (SST) over other time scales, e.g., Pliocene [Wycech et al., 2020].

At the same time, the predictors used in these models, the principal components of the temperature field at the proxy locations, are not exclusive and can be changed or extra information can be added as predictors, e.g. tree rings, cave data, ice cores, lake sediments, corals, etc. [Tingley and Huybers, 2010].

8 Future work

For the future work, besides tuning some of the conditions in the model, a sensitivity analysis with the number of proxy data can be performed and see how sensible the method is to the predictors presented.

Following the line presented in the conclusions, the future work lies in applying the BHM to other fields, such as rainfall, or vegetation indices over other domains, e.g.g India, South east Asia, etc. In this way, other features of the global climate can be better understood and more conclusions can be drawn from these future studies.

List of Figures

List of Tables

References

El niño and la niña. https://oceanservice.noaa.gov/facts/ninonina.html. Accessed: 2022-09-26.

El niño and la niña. https://www.unocha.org/themes/el-ni%C3%B1o/el-ni%C3%B1o-and-la-ni%C3%B1a#:~:text=During%20an%20El%20Ni%C3%B1o%20event,in%20temperature%2C%20rainfall%20and%20winds. Accessed: 2022-09-26.

C. Bracken, B. Rajagopalan, L. Cheng, W. Kleiber, and S. Gangopadhyay. Spatial bayesian hierarchical modeling of precipitation extremes over a large domain. *Water Resources Research*, 52(8):6643–6655, 2016. doi: https://doi.org/10.1002/2016WR018768. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016WR018768.

C. Bracken, K. D. Holman, B. Rajagopalan, and H. Moradkhani. A bayesian hierarchical approach to multivariate nonstationary hydrologic frequency analysis. *Water Resources Research*, 54(1):243–255, 2018. doi: https://doi.org/10.1002/2017WR020403. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017WR020403.

S. Brassell, G. Eglinton, I. Marlowe, U. Pflaumann, and M. Sarnthein. Molecular stratigraphy: A new tool for climatic assessment. *Nature*, 320:129–133, 03 1986. doi: 10.1038/320129a0.

K. M. Cobb, N. Westphal, H. R. Sayani, J. T. Watson, E. D. Lorenzo, H. Cheng, R. L. Edwards, and C. D. Charles. Highly variable el niño–southern oscillation throughout the holocene. *Science*, 339(6115):67–70, 2013. doi: 10.1126/science.1228246. URL https://www.science.org/doi/abs/10.1126/science.1228246.

F. Delage and S. Power. The impact of global warming and the el niño-southern oscillation on seasonal precipitation extremes in australia. *Climate Dynamics*, 54, 05 2020. doi: 10.1007/s00382-020-05235-0.

J. Emile-Geay, K. Cobb, M. Carré, P. Braconnot, J. Leloup, Y. Zhou, S. Harrison, T. Correge, H. Mcgregor, M. Collins, R. Driscoll, M. Elliot, B. Schneider, and A. Tudhope. Links between tropical pacific seasonal, interannual and orbital variability during the holocene. *Nature Geoscience*, 9, 12 2015. doi: 10.1038/ngeo2608.

M. K. Gagan, L. K. Ayliffe, D. Hopley, J. A. Cali, G. E. Mortimer, J. Chappell, M. T. McCulloch, and M. J. Head. Temperature and surface-ocean water balance of the mid-holocene tropical western pacific. *Science*, 279(5353):1014–1018, 1998. doi: 10.1126/science.279.5353.1014. URL https://www.science.org/doi/abs/10.1126/science.279.5353.1014.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992. ISSN 08834237. URL http://www.jstor.org/stable/2246093.

E. C. Gill, B. Rajagopalan, P. Molnar, and T. M. Marchitto. Reduced-dimension reconstruction of the equatorial pacific sst and zonal wind fields over the past 10,000 years using mg/ca and alkenone records. *Paleoceanography*, 31(7):928–952, 2016. doi: https://doi.org/10.1002/2016PA002948. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016PA002948.

E. C. Gill, B. Rajagopalan, P. H. Molnar, Y. Kushnir, and T. M. Marchitto. Reconstruction of indian summer monsoon winds and precipitation over the past 10,000 years using equatorial pacific sst proxy records. *Paleoceanography*, 32(2):195–216, 2017. doi: https://doi.org/10.1002/2016PA002971. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016PA002971.

T. Herbert. 6.15 - alkenone paleotemperature determinations. In H. D. Holland and K. K. Turekian, editors, *Treatise on Geochemistry*, pages 391–432. Pergamon, Oxford, 2003. ISBN 978-0-08-043751-4. doi: https://doi.org/10.1016/B0-08-043751-6/06115-6. URL https://www.sciencedirect.com/science/article/pii/B0080437516061156.

M. D. Hoffman and A. Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo, 2011. URL https://arxiv.org/abs/1111.4246.

D. W. Lea, T. A. Mashiotta, and H. J. Spero. Controls on magnesium and strontium uptake in planktonic foraminifera determined by live culturing. *Geochimica et Cosmochimica Acta*, 63(16):2369–2379, 1999. ISSN 0016-7037. doi: https://doi.org/10.1016/S0016-7037(99)00197-0. URL https://www.sciencedirect.com/science/article/pii/S0016703799001970.

G. Leduc, R. Schneider, J.-H. Kim, and G. Lohmann. Holocene and eemian sea surface temperature trends as revealed by alkenone and mg/ca paleothermometry. *Quaternary Science Reviews*, 29(7):989–1004, 2010. ISSN 0277-3791. doi: https://doi.org/10.1016/j.quascirev.2010.01.004. URL https://www.sciencedirect.com/science/article/pii/S0277379110000065.

C. R. Loader. Local likelihood density estimation. *The Annals of Statistics*, 24(4):1602 – 1618, 1996. doi: 10.1214/aos/1032298287. URL https://doi.org/10.1214/aos/1032298287.

M. M. Monteagudo, J. Lynch-Stieglitz, T. M. Marchitto, and M. W. Schmidt. Central equatorial pacific cooling during the last glacial maximum. *Geophysical Research Letters*, 48(3):e2020GL088592, 2021. doi: https://doi.org/10.1029/2020GL088592. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL088592. e2020GL088592 2020GL088592.

D. Nürnberg, J. Bijma, and C. Hemleben. Assessing the reliability of magnesium in foraminiferal calcite as a proxy for water mass temperatures. *Geochimica et Cosmochimica Acta*, 60(5):803–814, 1996. ISSN 0016-7037. doi: https://doi.org/10.1016/0016-7037(95)00446-7. URL https://www.sciencedirect.com/science/article/pii/0016703795004467.

A. Ossandón, B. Rajagopalan, U. Lall, J. S. Nanditha, and V. Mishra. A bayesian hierarchical network model for daily streamflow ensemble forecasting. *Water Resources Research*, 57(9):e2021WR029920, 2021. doi: https://doi.org/10.1029/2021WR029920. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021WR029920. e2021WR029920 2021WR029920.

B. Renard. A bayesian hierarchical approach to regional frequency analysis. *Water Resources Research*, 47(11), 2011. doi: https://doi.org/10.1029/2010WR010089. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010WR010089.

R. Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, and C. Yau. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1, 12 2021. doi: 10.1038/s43586-020-00001-2.

T. M. Smith, R. W. Reynolds, T. C. Peterson, and J. Lawrimore. Improvements to noaa's historical merged land–ocean surface temperature analysis (1880–2006). *Journal of Climate*, 21(10):2283 – 2296, 2008. doi: 10.1175/2007JCLI2100.1. URL https://journals.ametsoc.org/view/journals/clim/21/10/2007jcli2100.1.xml.

M. P. Tingley and P. Huybers. A bayesian algorithm for reconstructing climate anomalies in space and time. part i: Development and applications to paleoclimate reconstruction problems. *Journal of Climate*, 23(10):2759 – 2781, 2010. doi: 10.1175/2009JCLI3015.1. URL https://journals.ametsoc.org/view/journals/clim/23/10/2009jcli3015.1.xml.

J. B. Wycech, E. Gill, B. Rajagopalan, T. M. Marchitto Jr, and P. H. Molnar. Multiproxy reduced-dimension reconstruction of pliocene equatorial pacific sea surface temperatures. *Paleoceanography and Paleoclimatology*, 35(1):e2019PA003685, 2020. doi: https://doi.org/10.1029/2019PA003685. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019PA003685. e2019PA003685 2019PA003685.

Appendix

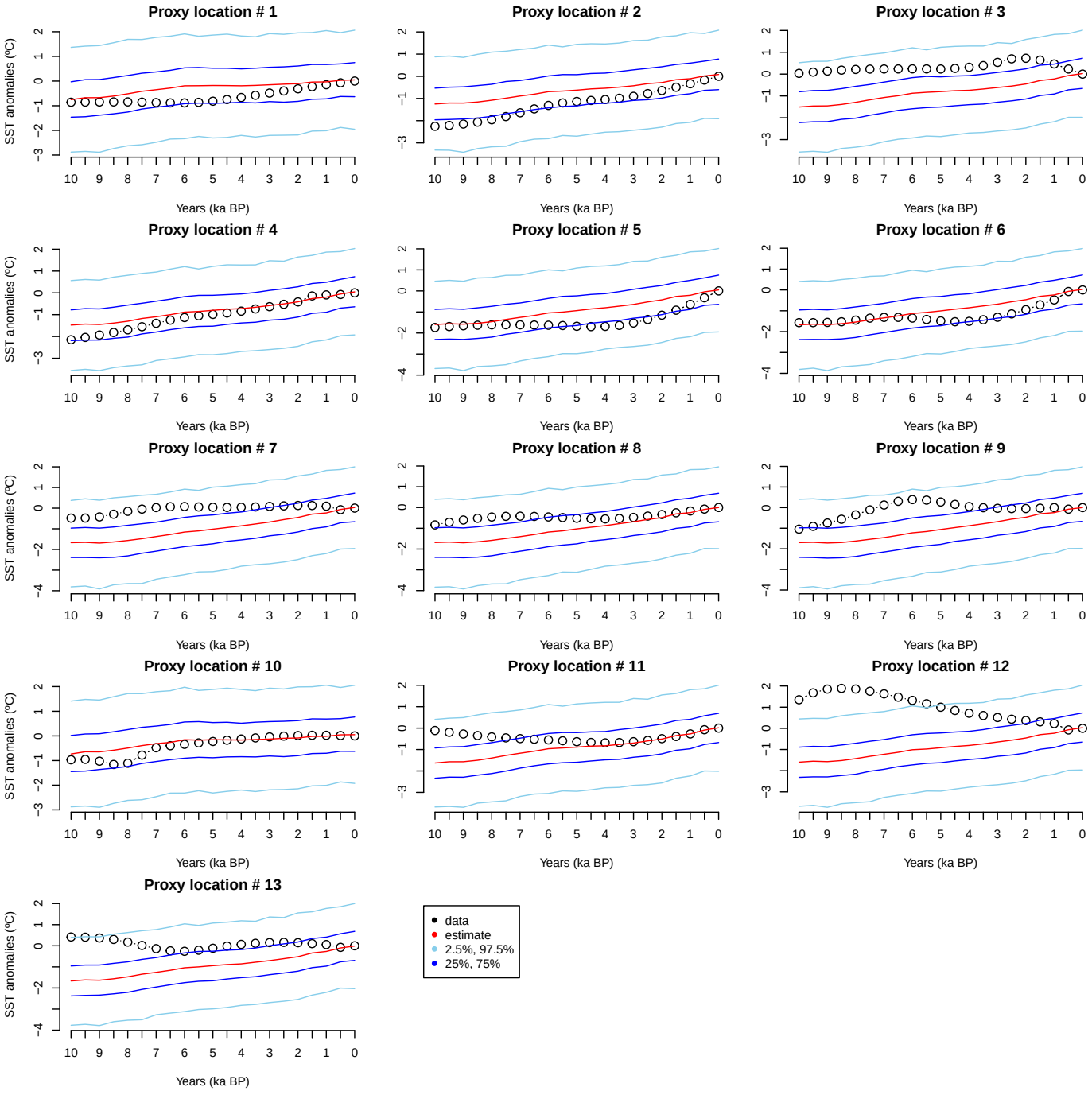

Figure 26: Scatterplots of each proxy record from the eastern Pacific (black) along with the reconstructed SST values for the grid point nearest each record (red) with the BHM.

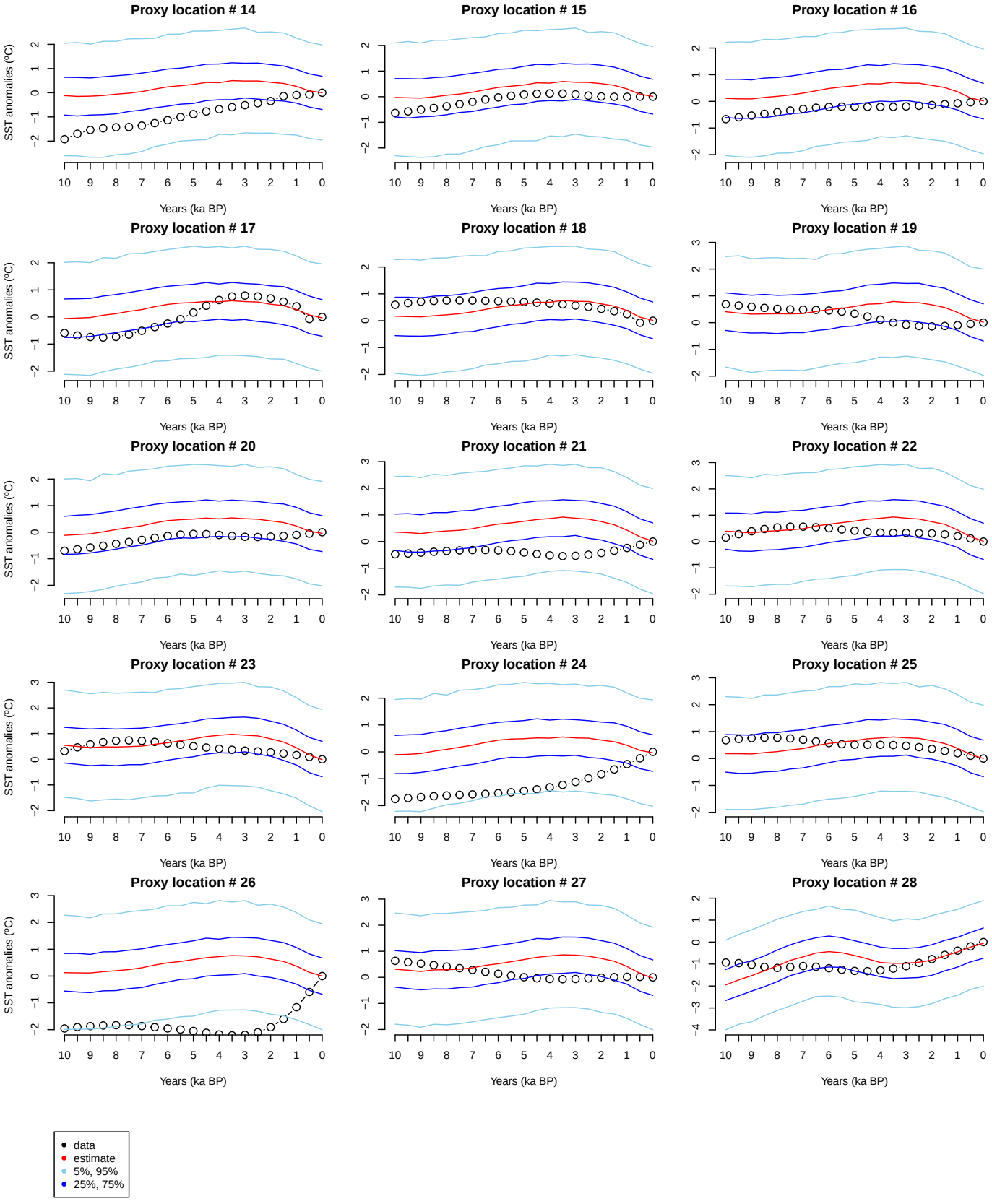

Figure 27: Scatterplots of each proxy record from the western and central Pacific (black) along with the reconstructed SST values for the grid point nearest each record (red) with the BHM.