HISEAS SST Reconstruction - RSOI Method

Anna Chen

Last Updated on January 30th, 2026

The **HISEAS** Project is an international paleo-climate research collaboration that aims to answer: "**How much and why did Ice Sheets melt during the Last Interglacial?**" My role in this project, thus far, has been focusing on quantitatively understanding the spatial and temporal patterns of Sea-Surface Temperature (SST) during the Last Interglacial (LIG, 130-115 Kya). This is done by reconstructing full-field SST maps across 130-115 Kya using the LIG SST dataset complied by Hoffman et al. (2017) [2].

This is a simplified run-down of preparing data matrices for the Reduced Space Optimal Interpolation (RSOI) method to reconstruct SST for the LIG.

1 Data Preparation

1.1 Construct Modern SST Data Matrix X

We use the Hadley Centre Sea Ice and Sea Surface Temperature dataset (HadISST) 1º-by-1º dataset [4] to construct a modern SST data matrix:

$$\mathbf{X} \in \mathbb{R}^{N \times T},$$

where N is the flattened full-field coordinate grids ($N = 180 \times 360 = 64800$) minus the locations with NaN values (see paragraph below). T is Age in years from 1870 to 2020 ($T = 151$). Each entry in $\mathbf{X}$ represents the SST of a given year in a given coordinate location.

We also need to remove rows with NaN data (e.g. land grids), but do not reset index. Later, after we have fully reconstructed the paleo SST field, $\mathcal{T}_t$, the reconstructed column vector $\mathcal{T}_t$ of time-step t should be indexed according to the indexes of the modern data matrix $\mathbf{X}$. We then fill in the missing rows with NaN. Therefore, when mapping, rows with NaN in the modern data matrix will also be marked as NaN in the reconstructed paleo SST data matrix. Hence, the number of rows, N, for the modern data matrix $\mathbf{X}$ is only around 60-70% the total number of grids (180×360), since the Earth is covered around 70% by ocean.

Note HadISST is a monthly gridded dataset. To account for a full ENSO cycle in each annual average, rather than averaging for the January-to-December calendar year, we would like to average from April of that given year to the next March. For example, the "1870" time-step actually would mean the average of 12 months from April 1870 to March 1871. Months with NaN values are ignored when averaging.

For the purpose of constraining Summer SST signals in the paleo dataset, which we will get to in Section 1.3, we should repeat this data matrix construction process for only the month of *August*, containing HadISST modern data of only August of the given time-step, also for a total of $T = 151$ time-steps. Remove rows with NaN data, similarly. Then, subtract each entry in the August-only matrix by each entry in the full modern data matrix $\mathbf{X}$ to compute an August-anomaly matrix. Call this anomaly matrix $\mathbf{A}$. Last, compute the row-wise mean of $\mathbf{A}$ to calculate, on average between 1870 and 2020, how much warmer or colder is the August SST compared to the annual SST at each location:

$$\mathbf{x}_{\text{Aug}} = \frac{\mathbf{A} \times \mathbf{1}_T}{T}$$

We will use the column vector $\mathbf{x}_{\text{Aug}}$ in Section 1.3 and 1.4.4.

1.2 Empirical Orthogonal Function (EOF) analysis on modern data matrix X

We should now centralize $\mathbf{X}$ over time, not space. This means we subtract each entry in $\mathbf{X}$ with the row-wise mean. Let $\mathbf{1}_T$ be a T-dimensional column vector. Then:

$$\overline{\mathbf{x}} = \frac{\mathbf{X} \times \mathbf{1}_T}{T}$$

and the centralized modern data matrix:

$$\mathbf{X}_{\text{Centralized}} = \mathbf{X} - \overline{\mathbf{x}} \times \mathbf{1}_T{}^{\mathsf{T}}$$

For all discussions below, we assume that we are working with the centralized data matrix, $\mathbf{X}_{\text{Centralized}}$, whenever we refer to $\mathbf{X}$. In Python, we can override $\mathbf{X}$ using $\mathbf{X}_{\text{Centralized}}$ and rename it as $\mathbf{X}$.

The (unbiased) covariance matrix of the centralized data matrix is therefore:

$$\mathbf{C} = \frac{1}{T-1}\mathbf{X}\mathbf{X}^{\mathsf{T}}$$

Conduct Singular Value Decomposition (SVD) on the centralized data matrix:

$$\mathbf{X} = \mathbf{E}\Sigma\mathbf{V}^{\mathsf{T}}$$

As described in Prof. Kaplan's email, the columns of $\mathbf{E}$ are non-dimensional spatial patterns, called *EOFs*. The columns of $\mathbf{V}$ (rows of $\mathbf{V}^\intercal$) are non-dimensional temporal patterns, called *amplitudes*. The Principal Components (PCs) are the rows of $\Sigma\mathbf{V}^\intercal$, or equivalently, the columns of $\mathbf{V}\Sigma$. PCs are dimensional.

The centralized data matrix is recovered as:

$$\begin{aligned}\mathbf{X} &= \mathbf{E}\Sigma\mathbf{V}^\intercal \\ &= \mathbf{E}(\underbrace{\mathbf{V}\Sigma}_{\text{PCs}})^\intercal\end{aligned}$$

Because EOF analysis decomposes the spatial covariance of the modern climate field into a set of orthogonal vectors and their corresponding variances, we may also represent the spatial covariance matrix as:

$$\mathbf{C} = \mathbf{E}\Lambda\mathbf{E}^\intercal$$

where Λ represents the diagonal matrix of eigenvalues of $\mathbf{C}$:

$$\Lambda = \frac{1}{T-1}\Sigma^2$$

We can perform a rank-20 approximation ($r = 20$), so we are left with $\mathbf{E}$ having 20 columns and $\mathbf{V}^\intercal$ having 20 rows. After rank reduction, the covariance matrix can be re-written using rank-reduced components as:

$$\mathbf{C} \approx \underset{N\times r}{\mathbf{E}}\underset{r\times r}{\Lambda}\underset{r\times N}{\mathbf{E}^\intercal}$$

We may plot the first 5, out of the total 20 columns of $\mathbf{E}$ and rows of $\mathbf{V}^\intercal$, to better visualize the temporal and spatial patterns during 1870-2020.

1.3 Area-Weighing for HadISST EOF and SST Matrix X

Recall that we have previously used Singular Value Decomposition (SVD) to decompose the centralized data matrix $\mathbf{X}$ as:

$$\begin{aligned}\mathbf{X} &= \mathbf{E}\Sigma\mathbf{V}^\intercal \\ &= \mathbf{E}(\underbrace{\mathbf{V}\Sigma}_{\text{PCs}})^\intercal\end{aligned}$$

where the columns of the orthogonal matrix $\mathbf{E}$ form an orthonormal basis of $\mathbb{R}$:

$$\mathbf{E}^\intercal\mathbf{E} = \mathbf{I}$$

and $\mathbf{I}$ represents an identity matrix.

It is, however, biased to assume that the columns of $\mathbf{E}$ (representing spatial patterns) should be orthonormal, because each spatial pattern are not weighed by the area of each grid they express. For example, a spatial pattern that describes polar regions would wrongly overemphasize the high-latitude patterns, because high-latitude grids have much smaller surface area than low-latitude grids.

Hence, we need to create a weight that weighs each entry of each column of $\mathbf{E}$ depending on the latitude of the grid, which is a cosine function. We may create an area-weighing factor like such:

$$w_i = \frac{\cos\varphi_i}{\sum_{j=1}^{N}\cos\varphi_j}, \qquad i = 1, \ldots, N$$

where φ refers to the latitude of the given SST grid.

Then, let us define the weight matrix as

$$\mathbf{W} = \mathrm{diag}(w_i)$$

where $\mathbf{W} \in \mathbb{R}^{N\times N}$.

Then, instead of having $\mathbf{E}$ as an orthogonal matrix, we should apply the weight as such:

$$\mathbf{E}^{\intercal}\mathbf{W}\mathbf{E} = \mathbf{I}$$

such that the area-weighed spatial pattern, $\mathbf{E}'$, is orthogonal:

$$\mathbf{E}' = \mathbf{W}^{\frac{1}{2}}\mathbf{E}$$

$$(\mathbf{E}')^{\intercal}\mathbf{E}' = \mathbf{I}$$

Now, recall from above that we have defined Principal Components (PCs) as $\mathbf{V}\Sigma$. We need to replace the area-biased $\mathbf{E}$ with the area-weighted $\mathbf{E}'$ into the previous SVD results. Also, we need to weight each SST value in $\mathbf{X}$ by the same weight, $\mathbf{W}$. So, we replace $\mathbf{X}$ with area-weighed $\mathbf{X}' = \mathbf{W}^{\frac{1}{2}}\mathbf{X}$ for SVD:

$$\begin{aligned}\mathbf{X} &= \mathbf{E}\Sigma\mathbf{V}^{\intercal}\\ \underbrace{\mathbf{W}^{\frac{1}{2}}\mathbf{X}}_{\mathbf{X}'} &= \underbrace{\mathbf{W}^{\frac{1}{2}}\mathbf{E}}_{\mathbf{E}'}\Sigma\mathbf{V}^{\intercal}\end{aligned}$$

And therefore we deduce the relationship between PCs and the area-weighed EOFs and modern SST data matrix:

$$\begin{aligned}\mathbf{X}' &= \mathbf{E}'\Sigma^{\intercal}\mathbf{V}\\ \mathbf{X}' &= \mathbf{E}'(\mathbf{V}\Sigma)^{\intercal}\\ (\mathbf{E}')^{\intercal}\mathbf{X}' &= (\mathrm{PC})^{\intercal}\end{aligned}$$

Finally, re-compute SVD using $\mathbf{X}'$ and $\mathbf{E}'$, and use those PCs computed using the equation above.

From this point onwards, we should remember to replace all $\mathbf{E}$ and $\mathbf{X}$ as their area-weighted counterparts, $\mathbf{E}'$ and $\mathbf{X}'$, respectively. Note that the notation below is written after this chunk is written, so all steps below still uses $\mathbf{E}$ and $\mathbf{X}$. But note that we should override these area-biased variables using the area-weighted variables.

Recall the August anomaly column vector $\mathbf{x}_{\text{Aug}}$ we used to extract the anomaly of August SST from the annual moderm data matrix $\mathbf{X}$. Because we have conducted area-weighing on $\mathbf{X}$, we also need to weight the August anomaly. Hence, we shall compute $\mathbf{x}'_{\text{Aug}}$ by scaling each i-th entry in $\mathbf{x}_{\text{Aug}}$ using $w_i^{\frac{1}{2}}$. Replace all $\mathbf{x}_{\text{Aug}}$ appearing below as the area-weighed $\mathbf{x}'_{\text{Aug}}$.

When we finally finish with the full-field SST reconstruction, we need to recover the physical temperature by reverting the results back using the weight. This will be explained in the final step.

1.4 Construct Paleo SST Data $\mathbf{y}_t$

1.4.1 Loading the Paleo SST Spreadsheet

The LIG SST dataset provided by Hoffman et al. (2017) [2], in an Excel spreadsheet format, was complied using various paleo-proxies: Foraminifera, Coccolithophores, UK37, Mg/Ca, Diatoms, and Radiolaria. The dataset includes n=104 pieces of SST data over 130-115 Kya (time-step = 0.1 Kya). Each sheet represents the time-series of SST data of the given location and proxy type.

- The first row is the header row.
  - Python command: skiprows=1, header=None
- The sheet name is the core's *ID*.
- Column B (cell B2) is *longitude*.
- Column C (cell C2) is *latitude*.
- Column R (cells R2 through R152) is *age*, or time-step, from 115.0 to 130.0 Kya. This is identical throughout all sheets.
- Column T (cells T2 through T152) is the paleo *SST* time-series of the given location.
- Column U (cells U2 through U152) contains the "SST + 2 standard deviation (δ)" data

- Column S, likewise, contains "SST - 2δ".

The (lon, lat) coordinates are exact to the hundredth place. We need to round them to the nearest 1ºx1º grid. We may use the following function: `lon_round = (np.floor(Original_longitude) + 0.5)`. Latitude likewise (`lat_round`). Then, we can build a 3-column dataframe with 104 cores' ID, rounded longitude, and rounded latitude data. Call this `df_coords`.

1.4.2 Optimal Estimation for Multi-proxy Cores

Note that some sheets' SST data are generalized based on the same core, but using different proxy types, resulting in "Multi-proxy cores". Some other cores are so close to each other that they are within the same 1ºx1º grid. Then, the estimated paleo SST will not match across the different proxy types, or multiple cores will be attributed to the same location. To resolve this conflict, rather than only using one core's SST and discarding the remaining core's data like we've done in the past, we use a technique called **Optimal Estimation**.

Say we use column U minus column T and divide that by 2 to derive 1δ for any given location and given time-step.

Let a random variable $\mathcal{Y}_t$ represent the paleo SST value Y_i and the standard deviation δ_i for each i-th proxy of a given location and a given time t. i should be no larger than 3 for each random variable, because there are no more than 3 overlapping cores for any given location. From here, we may calculate a weight (note that this weight shall not be confused with the area-weighing weight w_i from Step 1.3):

$$\mathcal{W}_i = \frac{\frac{1}{\delta_i^2}}{\sum \frac{1}{\delta_i^2}}$$

so that the optimally estimated SST of that location at the given time is:

$$\hat{\mathcal{Y}}_t = \sum \mathcal{W}_i Y_i$$

In total, there are 104 cores but only 82 unique coordinate pairs.

We repeat this method for each location and each time-step to build a paleo SST data column vector $\mathbf{y}_t$ for each time-step t:

$$\mathbf{y}_t \in \mathbb{R}^{n_t},$$

where n_t is the number of paleo-proxy data pieces contain valid values at time t. At most, there are 82 of these locations at any given t, so $n_t \leq 82$. t is Age in Kya (thousand of years before present) from 115.0 to 130.0 ($t = 151$).

Realize that, at some time-steps, not all 82 paleo-proxies contain valid SST values. After checking all sheets through Python, we find that, out of the 82 paleo-proxy sites we are using to construct the paleo SST data matrix, 2 of them do not contain the full 151 time-steps. Namely:

- Sheet-name **V28-238** at {lon, lat} = {160.48, 1.02} has 89 valid time-steps, starting 121.2 and ending 130.0 Kya.
- Sheet-name **RC11-86_Foram** at {lon, lat} = {18.45, -35.78} has 150 valid time-steps, only missing the 115.0 Kya time-step.

Therefore, the number of available paleo-proxies n_t at different time-steps $t \in \{115.0, 115.1, \ldots 130.0\}$ can be represented as:

$$n_t = \begin{cases} 80 & : \quad t = 115.0 \\ 81 & : \quad 115.1 \leq t \leq 121.1 \\ 82 & : \quad t \geq 121.2 \end{cases}$$

1.4.3 Area-Weighing for Paleo SST $\mathbf{y}_t$

Recall that we have previously applyed an area-weight to all EOFs and the modern SST data $\mathbf{X}$. We need to do the same to all paleo SST data, $\mathbf{y}_t$ as well. Hence, using the same weight function as used for $\mathbf{E}'$ and $\mathbf{X}'$:

$$w_i = \frac{\cos \varphi_i}{\sum_{j=1}^{n_t} \cos \varphi_j}$$

compute a weight w_i for each i-th paleo-proxy SST based on the latitude φ_i of the given proxy. Then, we may scale each i-th paleo-SST in $\mathbf{y}_t$ by $w_i^{\frac{1}{2}}$ for each time-step t. Hence, from now on, all variables $\mathbf{y}_t$, $\mathbf{x}_{\text{Aug}}$, $\mathbf{X}$, and $\mathbf{E}$ are area-weighed even if they do not have the apostrophe ($'$) sign.

1.4.4 Correcting Summer Signal Biases

An issue with Hoffman et al.'s dataset is that some paleo-proxies are inferred to display *summer* (August) signals, whereas most others are inferred to show *annual* signals. The methodology of Hoffman et al. to calculate annual signals was simply to average the Summer (August) and Winter (February) SST estimates. This is perhaps less ideal than what we'd expect. However, to work with the discrepancies in signal seasonality, we need to do the following:

1. Identify, from Table S1 in Hoffman et al.'s supplements, the IDs of summer signal proxies.
   - Sheet names of the summer signal cores: [”M23323-1_UK37”, ”EW9302-JPC8”, ”ODP 980”, ”NA87-25”, ”NEAP18k”, ”MD04-2845”, ”SU92-03”, ”CH69-K09”, ”MD95-2040_Foram”, ”SU90-03”, ”MD95-2036”, ”KNR140-37JPC”, ”ODP 1089_Radiolaria”, ”MD94-101”, ”PS2489-2”, ”MD84-527”, ”MD94-102”, ”MD88-770_Diatom”, ”MD02-2488”]
2. Identify the corresponding coordinate locations (`lon_round` and `lat_round`) of these IDs.

3. Identify the rows in each $\mathbf{y}_t$ and $\mathbf{X}$ in which such ID corresponds to.

4. Realize that $\mathbf{X}$ and the August anomaly column vector $\mathbf{x}_{\text{Aug}}$ have the same number of rows. Hence, for all rows in $\mathbf{x}_{\text{Aug}}$ that are *not* summer-signal rows in $\mathbf{X}$ found in the previous step, change the value of the row's entry to 0.

5. For each row in $\mathbf{y}_t$ that correspond to each row in $\mathbf{x}_{\text{Aug}}$, subtract the entry in $\mathbf{y}_t$ by the entry in $\mathbf{x}_{\text{Aug}}$, and repeat for each time-step t. This means we are subtracting the August anomaly from each location that describes a summer proxy signal, manipulating the corresponding row show a paleo SST value closer to an annual signal.

Note that a big assumption behind this method is that the summer-annual anomaly is stationary from 115 to 130 Kya, and that the scale of the SST anomaly is the same from modern (1870-2020) to paleo (115-130 Kya) periods.

We build a matrix $\mathbf{H}$ as a submatrix of the $N \times N$ identity matrix only containing rows corresponding to the n_t locations where paleo SST data is available for that given time-step:

$$\mathbf{H}_t \in \mathbb{R}^{n_t \times N}$$

where N refers to the full-field modern data matrix's number of rows, i.e. 180×360 minus the number of NaN rows. Because n_t is dependent on t, $\mathbf{H}_t$ is also dependent on t. We can use this $\mathbf{H}_t$ to select n_t paleo-proxy rows of a full-field SST matrix.

After correcting rows previously displaying summer signals in $\mathbf{y}_t$, we can centralize the corrected $\mathbf{y}_t$, like we've done for $\mathbf{X}$, using the row-wise mean of the modern data matrix: $\overline{\mathbf{x}}$.

$$\mathbf{y}_{\text{Centralized, } t} = \mathbf{y}_t - \mathbf{H}_t \overline{\mathbf{x}}$$

Like how we've overrode $\mathbf{X}$ with $\mathbf{X}_{\text{Centralized}}$, we would also override each column vector $\mathbf{y}_t$ with $\mathbf{y}_{\text{Centralized, } t}$ and still refer to the centralized column vector as $\mathbf{y}_t$ for simplification.

2 Reduced Space Optimal Interpolation (RSOI)

1. Let us define that, at each time t:

$$\underset{n_t \times r}{\mathcal{H}_t} = \underset{n_t \times N}{\mathbf{H}_t} \underset{N \times r}{\mathbf{E}}$$

where $r = 20$ is the number of ranks we wish to truncate $\mathbf{E}$, and n_t is the available number of proxies at that time-step t. Note that $\mathbf{H}_t$ is a function of t, but $\mathbf{E}$ is not.

Then, we can find a column vector $\mathbf{p}_t$ at time t where:

$$\begin{aligned}
\underset{r\times 1}{\mathbf{p}_t} &= \underbrace{\Lambda\mathcal{H}_t^\intercal}_{r\times n_t}(\underbrace{\mathcal{H}_t\Lambda\mathcal{H}_t^\intercal}_{n_t\times n_t} + \underset{n_t\times n_t}{\mathbf{R}_t})^{-1}\underset{n_t\times 1}{\mathbf{y}_t} \\
&= \underbrace{(\Lambda^{-1} + \mathcal{H}_t\mathbf{R}^{-1}\mathcal{H}_t^\intercal)^{-1}}_{\text{Error in } \mathbf{p}_t}\mathcal{H}_t\mathbf{R}^{-1}\mathbf{y}_t
\end{aligned}$$

which we can prove to be equivalent. To explain the variables above:

- First recall what we had from SVD on modern data:

$$\underset{N\times T}{\mathbf{X}} = \underset{N\times T}{\mathbf{E}} \quad \underset{T\times T}{\Sigma} \quad \underset{T\times T}{\mathbf{V}^\intercal}$$

- Λ is the diagonal matrix of eigenvalues of modern data's spatial covariance matrix ($\mathbf{C}$), calculated earlier through EOF, and reduced to rank r=20.

- $\mathbf{y}_t$ is the column vector of paleo SSTs at time t with dimension n_t, and $n_t \leq 82$. Note that $\mathbf{y}_t$ has already been area-weighed by $\mathbf{W}^{\frac{1}{2}}$.

- $\mathbf{R}_t$ is the error matrix, elaborated in the next point.

Now that we have obtained $\mathbf{p}_t$ for each time-step t, we may reconstruct the full-field SST as:

$$\mathcal{T}_t = \mathbf{E}\mathbf{p}_t$$

where $\{\mathcal{T}_t\}$ represent a set of gridded fields of $\mathcal{T}$, which are re-arranged as column vectors of dimension N, in order of ascending (from 115.0 to 130.0 Kya) time $t \in \{1, 2, \ldots T\}$.

2. It is important to account for error in SST estimation. How can we calculate $\mathbf{R}_t$? First, understand that we may decompose $\mathbf{R}_t$ into two components:

   (a) Observational error, $\mathcal{E}_t^{obs}$, from proxy-to-SST transfer function, e.g. error in the measurement of $d^{18}O$ from Forams, which was then computed to an estimated paleo SST.

   (b) Rank reduction error, $\mathcal{E}_t^{red}$, from the earlier SVD process, where the full-rank $\mathbf{E} \in \mathbb{R}^{N\times T}$ was truncated to 20-rank, $\mathbf{E} \in \mathbb{R}^{N\times r}$, $r = 20$.

   Mathematically, we may write the total error, $\mathbf{R}_t$, as the sum of the expected value of the two error components, and apply the linearity of expectations:

$$\begin{aligned}
\mathbf{R}_t &= \mathbb{E}(\mathcal{E}_t\mathcal{E}_t^\intercal) \\
&= \mathbb{E}(\mathcal{E}_t^{red}(\mathcal{E}_t^{red})^\intercal) + \mathbb{E}(\mathcal{E}_t^{obs}(\mathcal{E}_t^{obs})^\intercal)
\end{aligned}$$

Let us first find the rank reduction error, $\mathcal{E}_t^{red}$. Recall that we've always been using the truncated $\mathbf{E}$ with rank = 20. We can define a new variable, $\tilde{\mathbf{E}}$, from the remaining ranks from a total of 151 ranks, i.e. all columns of the original $\mathbf{E}$ except for the first 20 columns.

Likewise, we have truncated the diagonal matrix of eigenvalues Λ to the first 20 eigenvalues during SVD, and scaled by $\frac{1}{T-1}$. We can also define $\tilde{\Lambda}$ as the diagonal matrix of all remaining, unused eigenvalues, also scaled by $\frac{1}{T-1}$.

Therefore, we may compute a covariance matrix of the leftover ranks, then multiplied by the selector matrix $\mathbf{H}_t$ on both sides to select only the locations with paleo-proxies. This is equivalent to the rank reduction error component of the total error $\mathbf{R}_t$:

$$\mathbb{E}(\mathcal{E}^{red}(\mathcal{E}^{red})^\intercal) = \mathbf{H}_t(\tilde{\mathbf{E}}\tilde{\Lambda}\tilde{\mathbf{E}}^\intercal)\mathbf{H}_t^\intercal$$

Next, let us find the observational error, $\mathcal{E}_t^{\text{obs}}$. To find the deterministic matrix of this error, we can compute the diagonal matrix of variance for all available proxy locations at time t. We know that, for each sheet in the paleo SST Excel file, columns S and U contain "SST -2sd" and "SST +2sd" data for each time-step, respectively. Let σ represent standard deviation. Then, we may obtain σ by subtracting column T (the SST data column) from column U, and divide that by 2. Then, σ^2 is the variance. Thus, we can write a column vector $\mathbf{r}_{\text{obs}}$ for all n_t proxies for the given time-step:

$$\mathbf{r}_{\text{obs}} = \begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \vdots \\ \sigma_{n_t}^2 \end{bmatrix}$$

Then, weigh them by area in the same method as we've weighed $\mathbf{y}_t$, which is to multiply each $\mathbf{r}_{\text{obs, i}}$ by w_i previously computed for $\mathbf{y}_t$ above. We are using w_i rather than $w_i^{\frac{1}{2}}$ for the weight because r_{obs} is made up of variance, which is the square of σ, so equally the weight of each σ must be squared as well.

The observational error of a given time t is a diagonal one with components of $\mathbf{r}_{\text{obs}}$ on the diagonal:

$$\mathbb{E}(\mathcal{E}_t^{obs}(\mathcal{E}_t^{obs})^\intercal) = \text{diag}(\mathbf{r}_{\text{obs}})$$

Now that we have finished computing the two components of error (the sum of the two expected values is the total error, $\mathbf{R}_t$), how does the error relate to the reconstructed SST data? We may represent the sum of the two error components as the difference between the reconstructed SST and the given SST from paleo-proxies at the locations where paleo-proxies are provided.

Taking into account of the rank reduction error, the reconstructed full SST field should be written as:

$$\mathcal{T}_t = \mathbf{E}\mathbf{p}_t + \tilde{\mathbf{E}}\tilde{\mathbf{p}_t}$$

where the tilde sign represents the variables computed using the leftover ranks after the 20th rank.

If we select the locations from the reconstructed SST ($\mathcal{T}_t$) that matches the given paleo-proxy locations, we can write the selected SST values as $\mathbf{H}_t\mathcal{T}_t$. It is impossible for this reconstruction to be exactly the same as the given SST from paleo-proxies ($\mathbf{y}_t$), because there is a observational error innate to $\mathbf{y}_t$. Hence, we can express the small difference between the reconstruction and the given values as the sum of the two components of error:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{H}_t\mathcal{T}_t + \mathcal{E}_t^{obs} \\ &= \mathbf{H}_t\mathbf{E}\mathbf{p}_t + \underbrace{\mathbf{H}_t\tilde{\mathbf{E}}\tilde{\mathbf{p}_t}}_{\mathcal{E}^{red}} + \mathcal{E}_t^{obs} \end{aligned}$$

This is what the errors actually mean in terms of their impacts on the reconstructed SST values.

3. Recall that we have performed area-weighing in Step 1.3 ”Area-Weighing for HadISST EOF” above. This indicates that the obtained reconstructed full-field SST, $\mathcal{T}_t$, is also area-weighed, and needs to be un-weighted to represent the actual physical SST.

   We had from above that:

   $$\mathcal{T}_t = \mathbf{E}\mathbf{p}_t$$

   and note the spatial pattern $\mathbf{E}$ here denotes the area-weighted EOFs (recall we did $\mathbf{E}' = \mathbf{W}^{\frac{1}{2}}\mathbf{E}$, and then we used $\mathbf{E}$’ to override all subsequent $\mathbf{E}$). To ”un-weight” the resulting $\mathcal{T}$ so the physical SST can be obtained:

   $$\mathcal{T}_{t,\text{ phys}} = \mathbf{W}^{-\frac{1}{2}}\mathbf{E}\mathbf{p}_t$$

   Note that $\mathcal{T}_{t,\text{phys}}$ should be the final reconstructed results that we plot, because the area-weighed temperature values do not have any physical meaning.

4. (Note this chunk continues to use $\mathbf{E}$ and $\mathbf{X}$ to represent the area-weighed variables, because this chunk is written prior to performing area-weighing.)

Currently, when mapping the spatial and temporal patterns for our modern EOF, we saw mostly negative values for the leading rank. An optional change to make is to change the signs of PCs and EOFs so the leading rank(s) contains mostly non-negative values, making it more reasonable when presenting. Realize that during SVD, we had:

$$\begin{aligned}\mathbf{X} &= \mathbf{E}\Sigma\mathbf{V}^\intercal \\ &= \mathbf{E}(\mathbf{V}\Sigma)^\intercal \\ &= \mathbf{E}(\text{PC})^\intercal\end{aligned}$$

To change the signs of both $\mathbf{E}$ and PC, we can create a diagonal matrix $\mathbf{D}$ that only contains -1 on its diagonal entries, i.e. a negative identity matrix. Note that $\mathbf{D} = \mathbf{D}^{-1}$. Hence:

$$\begin{aligned}\mathbf{X} &= \mathbf{E}\Sigma\mathbf{V}^\intercal \\ &= (\mathbf{E}\mathbf{D})\Sigma(\mathbf{D}^{-1}\mathbf{V}^\intercal) \\ &= (\mathbf{E}\mathbf{D})\Sigma(\mathbf{V}\mathbf{D})^\intercal\end{aligned}$$

Therefore, we can update the matrix of spatial patterns ($\mathbf{E}$) and the PCs accordingly.

5. A large assumption that should be challenged is that the eigenvalues of the covariance matrices, Λ, for the Last Interglacial data and modern data are the same. This assumption is especially contestable if our paleo SST is significantly different from the modern data. For example, we know that the last glacial maximum was 6^{o}C colder than today's average temperature. How should we know if we should not constrain $\mathbf{p}_t$ to the same covariance as Λ?

   To check for this, we should examine the consistency between the full variance of signal versus the sum of the variance and the expectation of error:

$$\Lambda = \mathbb{E}(\mathbf{p}_t\mathbf{p}_t^\intercal) + \mathcal{P}_t^{\text{OI}}$$

   where:

$$\mathbf{P}_t = \mathbf{E}\mathcal{P}_t^{\text{OI}}\mathbf{E}^\intercal$$

   and the error of paleo SST estimate is:

$$\mathcal{P}_t^{\text{OI}} = \Lambda - \Lambda\mathcal{H}_t^\intercal(\mathcal{H}_t\Lambda\mathcal{H}_t^\intercal + \mathbf{R}_t)^{-1}\mathcal{H}_t\Lambda$$

   The meaning of this is that if we don't have any paleo data, we would have $\mathbf{p}_t = 0$ and the error in estimate $\mathbf{p}_t^{\text{OI}} = \Lambda$. If the error of SST estimate is severely inconsistent (having large discrepancies on the left and right sides?), then we should challenge our assumption.

   How to visualize error: (updated on Jan. 30th)

(a) Construct the full-field error

$$\mathbf{P}_t = \mathbf{E}\mathcal{P}_t^{\mathrm{OI}}\mathbf{E}^\intercal$$

using the equations above. This should be a $n_t \times n_t$ matrix.

(b) Extract the diagonal of $\mathbf{P}_t$ at each time t, and take the square root of each entry. This is the standard deviation of the error.

$$\text{Error SD} = \sqrt{\mathrm{diag}(\mathbf{P}_t)}$$

There should be n_t entries, each representing one grid on the global field, successfully forming a spatial pattern of error SD at each time t.

6. One optional idea of further improvement, not necessary at this step and currently not implemented, is to use modeled SST to fill in temporal gaps. Currently, our temporal resolution of paleo SST is every 0.1 Kya (i.e. every century). This presents a significant mismatch with the resolution of 1 Yr in the modern SST dataset. Hence, some patterns that cannot be captured during a time-frame as short as 1870-2020 may play a large role during the Last Interglacial, such as obliquity forcing.

   However, if we can use modeled products to interpolate between time-slices and increase the temporal resolution, we can perhaps work with the same short modern time-frame (1870-2020), because we assume short-term SST patterns can be captured with the paleo model.

2.1 RSOI Summary: Steps to Take

In conclusion, after understanding how each variable is derived from the steps listed above, we can reconstruct a full-field SST map $\mathcal{T}_t$ for each time-step t by doing the following:

1. Construct the error matrix $\mathbf{R}_t$ by:

$$\mathbf{R}_t = \mathbb{E}(\mathcal{E}_t^{red}(\mathcal{E}_t^{red})^\intercal) + \mathbb{E}(\mathcal{E}_t^{obs}(\mathcal{E}_t^{obs})^\intercal)$$

   where:

$$\mathbb{E}(\mathcal{E}_t^{obs}(\mathcal{E}_t^{obs})^\intercal) = \mathrm{diag}(\begin{bmatrix}\sigma_1^2 & \sigma_2^2 & \dots & \sigma_{n_t}^2\end{bmatrix}^\intercal)$$

$$\mathbb{E}(\mathcal{E}_t^{red}(\mathcal{E}_t^{red})^\intercal) = \mathbf{H}_t\tilde{\mathbf{E}}\tilde{\mathbf{p}}_t$$

2. Construct the time-variant column vector $\mathbf{p}_t$ by

$$\mathbf{p}_t = \Lambda\mathcal{H}_t^\intercal(\mathcal{H}_t\Lambda\mathcal{H}_t^\intercal + \mathbf{R}_t)^{-1}\mathbf{y}_t$$

3. Construct the full-field SST map as a column vector $\mathcal{T}_t$ by (remember $\mathbf{E}$ here should already be area-weighed using $\mathbf{W}^{\frac{1}{2}}$ previously):

$$\mathcal{T}_t = \mathbf{E}\mathbf{p}_t$$

4. Undo area-weighing of $\mathcal{T}_t$ and obtain SST reconstruction with actual physical meaning by:

$$\mathcal{T}_{t,\ \text{phys}} = \mathbf{W}^{-\frac{1}{2}}\mathbf{E}\mathbf{p}_t$$

5. Correspond each row index in $\mathcal{T}_{t,\ \text{phys}}$ to a coordinate on the 180 × 360 field to map the value of SST to that grid.

6. Repeat the previous 4 steps at all time-steps from 115.0 to 130.0 Kya to produce 151 reconstruction maps. Create an animation of these maps, starting with 130.0 and ending with 115.0 to represent the correct direction of time's arrow.

7. Visualize the error standard deviation $\mathbf{P}_t$ at each time-step as a global-field animation across 115.0-130.0 Kya.

References

[1] Evans, M. N., A. Kaplan, and M.A. Cane, 2002: Pacific sea surface temperature field reconstruction from coral d18O data using reduced space objective analysis,*Paleoceanography*, **17**(1), 10.1029/2000PA000590.

[2] Hoffman, J. S., Clark, P. U., Parnell, A. C., & He, F. (2017). Regional and global sea-surface temperatures during the last interglaciation. Science, 355(6322), 276–279. https://doi.org/10.1126/science.aai8464

[3] Kaplan, A., Y. Kushnir, M. Cane, and M. Blumenthal, 1997: Reduced space optimal analysis for historical datasets: 136 years of Atlantic sea surface temperatures,*J. Geophys. Res.*, **102**, 27835-27860.

[4] Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., et al. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. Journal of Geophysical Research: Atmospheres, 108(D14). https://doi.org/10.1029/2002JD002670