Paleoclimate Sampling as a Sensor Placement Problem

MAUD COMBOUL AND JULIEN EMILE-GEAY
University of Southern California, Los Angeles, California

GREGORY J. HAKIM
Department of Atmospheric Sciences, University of Washington, Seattle, Washington

MICHAEL N. EVANS
Department of Geology and Earth System Science Interdisciplinary Center, University of Maryland, College Park, College Park, Maryland

(Manuscript received 25 November 2014, in final form 3 July 2015)

ABSTRACT

This study formulates the design of optimal observing networks for past surface climate conditions as the solution to a data assimilation problem, given a realistic proxy system model (PSM), paleoclimate observational uncertainties, and a network of current and proposed observing sites. The method is illustrated with the design of optimal networks of coral $\delta^{18}O$ records, chosen among candidate sites, and used to jointly infer sea surface temperature (SST) and sea surface salinity (SSS) fields from the Community Climate System Model, version 4, last millennium simulation over the 1850–2005 period. It is shown that an existing paleo-observing network accounts for approximately 20% of the SST variance, and that adding 25 to 100 optimal pseudocoral sites would boost this fraction to 35%–52%. Characterizing the SST variance alone, or jointly with the SSS, leads to similar optimal networks, which justifies using coral $\delta^{18}O$ records for SST reconstructions. In contrast, the network design for reconstructing SSS alone is fundamentally different, emphasizing the hydroclimatic centers of action of El Niño–Southern Oscillation. In all cases, network design depends strongly on the amplitude of the observational error, so replicates may be more beneficial than the exploration of new sites; these replicates tend to be chosen where proxies are already informative of the large-scale climate field(s). Finally, extensions to other types of paleoclimatic observations are discussed, and a path to operationalization is outlined.

1. Introduction

A general goal of paleoclimatology is to use paleoclimate observations, which are sparser and noisier than direct observations, to characterize preinstrumental climate variability. Paleoclimate observations involve, for example, using $\delta^{18}O$ of coral aragonite to measure a combination of temperature during calcification and seawater $\delta^{18}O$ value, while direct observations would be made, for instance, by using a thermometer to measure bucket SST. Ideally, the design of sampling strategies includes both local and global considerations in some objectively determined combination. Local considerations include the presence and availability of instrumental records and material archives that may be sampled to produce paleoclimate observations, the sensitivity of the proxy system to climate conditions at the site, accessibility, safety, and costs. For scientific goals that include the description of nonlocal phenomena, additional considerations include the extent to which site observations reflect large-scale patterns, the variables in which those patterns are locally expressed, and the balance between expanding the observing network spatially and/or temporally and replicating existing data to reduce paleoclimate observational uncertainties. Although the former category of network design criteria is highly site specific, the latter is approachable by means of hypothetical studies. Given assumptions about local optimization criteria, how would one prioritize sampling

Corresponding author address: Julien Emile-Geay, Dept. of Earth Sciences, University of Southern California, 3651 Trousdale Parkway, Los Angeles, CA 90089.
E-mail: julieneg@usc.edu

DOI: 10.1175/JCLI-D-14-00802.1

© 2015 American Meteorological Society
expeditions given limited resources? Would it be more efficient to generate replicates of a proxy series at an existing site or go sample a new one? How much would each paleoclimate observation contribute to reducing uncertainty in our knowledge of past climates? Similar questions were explored by Bradley (1996), who considered an existing proxy network and examined how well it predicted global mean annual temperature series. Evans et al. (1998b) explored the more general problem of optimal proxy sampling for minimizing theoretical error in reconstructed fields, but for the more specific estimation of the SST field. To that end, the authors used a reduced-space optimal interpolation (OI)-based reconstruction framework (Gandin 1965; Kaplan et al. 1997) and assumed that hypothetical SST observations made with random error might be available at a predetermined, quasi-realistic, coral-atoll-like sampling network. For small paleoclimate observational errors, they found that relatively small networks were able to resolve the most important large-scale patterns of SST variance (e.g., Evans et al. 1998b, their Figs. 4–6) and reduce the analysis error by up to 30% regionally. In the presence of relatively large paleoclimate observational errors, re-sampling (replication) was often preferred over sampling additional locations. Their iterative, “next best” approach did not, however, guarantee that the resulting proxy network would be optimal; nor did they consider that the most common paleoclimatic observation made in coral aragonite is oxygen isotopic composition ($\delta^{18}O$), which depends not only on temperature but also seawater $\delta^{18}O$ at time of calcification (Evans et al. 1998b).

Recently, Mauger et al. (2013) solved the related problem of designing optimal climatological networks for a specified field using a related data assimilation framework (Ansell and Hakim 2007; Hakim and Torn 2008). They applied their method to monitoring two separate quantities: U.S. Pacific Northwest regionally and annually averaged precipitation and temperature. By minimizing the error about the mean temperature field, they found that a relatively small number of stations (5–10) was sufficient to characterize the mean-field variance over the domain. This approach, generalized to vector optimization (N. Hryniew and G. J. Hakim 2015, unpublished manuscript) and applied to paleoclimate observations mapped from multivariate fields, provides a solid foundation for the present work.

It turns out that such problems fall under the general umbrella of optimal sensor placement (OSP), which is the subject of an extensive literature. Roughly speaking, OSP problems can be divided into two main categories:

1) monitoring for event detection, for instance in large-scale urban infrastructures (see Berry et al. 2005; Krause et al. 2008a; Hart and Murray 2010; Comboul and Ghanem 2013), and

2) using sparse sensor networks to estimate a physical process in space and time (see kriging and its generalization to Gaussian process regression; e.g., Zimmerman 2006; Krause et al. 2008b; Castello et al. 2010).

In this study we formalize the design of optimal strategies for climate proxy sampling by borrowing tools from this literature. We generalize previous approaches so the paleo observational network design is solely based on the choice of information utility measure and the ability of each proxy to capture environmental signals. Three ingredients make this approach feasible: 1) an efficient formulation of the sensitivity problem (Ansell and Hakim 2007; Hakim and Torn 2008; Mauger et al. 2013); 2) models relating climate fields to paleoclimate observations, that is, proxy system models (PSMs; Evans et al. 2013); and 3) an efficient search algorithm (e.g., greedy algorithm; Nemhauser et al. 1978; Krause et al. 2008b; Das and Kempe 2008, 2011) if a direct search is not feasible. Given a set of possible sampling locations dictated by a set of presumed site-local biological and geological considerations, we seek to characterize the variance of one or more climate fields with a minimum number of paleoclimate observations of a given quality. To illustrate this process, we focus on designing an optimal network of oxygen isotope observations in coral aragonite ($\delta^{18}O_c$), among a list of possible observing sites, with the goal of learning about sea surface temperatures (SST) and/or sea surface salinity (SSS) from the Community Climate System Model, version 4 (CCSM4), last millennium simulation of Landrum et al. (2013). More precisely, given $p_0$ locations where coral $\delta^{18}O_c$ could potentially be measured, where are the best $p$ locations to capture as much variance in SST and SSS as possible? To do so, we leverage the simple bivariate model of Thompson et al. (2011) with a perfect model scenario aimed at establishing a proof of concept. However, the framework is extremely general, and applicable to any proxy system that is well understood enough to be numerically modeled via a PSM. A summary of the methodology is provided in Fig. 1, which illustrates how the different components of our approach operate together. The article is structured as follows: in section 2 we formulate the mathematical problem and propose a solution method. A pseudocoral example is introduced in section 3, which describes the setting and experimental design. Results are shown in section 4, followed by a discussion in section 5. We conclude with an assessment of challenges that remain before this strategy can be operationalized.

2. Methodology

Let $V$ denote the set of $p_0$ paleoclimate observation sites (i.e., $|V| = p_0$ is the cardinal of $V$). To $V$, we
associate a set of observable random variables (RV) \( y \). In our case, \( y \) represents the proxy measurement anomalies such as coral \( \delta^{18}O \) or strontium and/or calcium taken at specific sites. We denote by \( F \), where \( |F| = q \), the set of grid points and \( x \) a climate field at those locations, observed indirectly via \( y \). In the following, \( x \) will stand for anomalies of sea surface temperature and/or salinity and \( J \) for the field information we wish to infer from the proxies (e.g., the whole SST field, the Niño-3.4 index, or some other metric derived from SST and/or SSS), also called the forecast metric or response function of \( x \). Next, we describe the Gaussian model representation for \( x \), which is suitable for field anomaly values (centered around zero), and the Bayesian estimation of the system given paleoclimate observations \( y_S \) at some location \( S \).

### a. Sensor placement as a filtering problem

Let us consider the filtering problem consisting of a Gaussian process \( x \), which we cannot observe directly, and an observation process \( y_S \) correlated with \( x \) as follows:

\[
x \sim \mathcal{N}(\mu, \Sigma) \quad \text{and} \quad y_S = H_S(x) + e_S,
\]

where \( \mu \) and \( \Sigma \) are the prior mean and error covariance matrix of \( x \), \( y_S \) is the vector of \( p \) observations at a subset of locations \( S \in V \), \( H_S \) is the operator mapping the field variables to the measurements at \( S \). Since \( x \) is typically not directly observable (i.e., \( y_S \neq x_S + e_S \)), the operator is derived from a forward model such as the one described in section 3b and is generally nonlinear. In addition, \( e_S \sim \mathcal{N}(0, R_S) \) is the error vector, which includes measurement noise and the portion of the proxy archive that is not explained by the PSM, where \( R \) is a \( p_0 \times p_0 \) diagonal matrix (assuming the errors are independent) and its submatrix restricted to \( S \) is denoted by \( R_S \).

Then, we can estimate the posterior distribution of \( x \) given these observations \( y_S \). It is also Gaussian with conditional mean \( \mu^* \) and covariance \( \Sigma^* \) derived via Bayes’s theorem (e.g., Wikle and Berliner 2007):

\[
\mu^* = \mu + \Sigma H_S^T (H_S \Sigma H_S^T + R)^{-1} [y_S - H_S \mu] \quad \text{and} \quad \Sigma^* = \Sigma - \Sigma H_S^T (H_S \Sigma H_S^T + R)^{-1} H_S \Sigma.
\]

Here \( H_S \) is the first-order linearization of the observation operator \( H_S \). Placing this problem in context, kriging (or equivalently Gaussian process regression) is the case where \( H_S \) is just the identity matrix at sites \( S \) applied to \( x \) (in other words, \( H_S \) has ones on the diagonal indices corresponding to sites \( S \) and zeros elsewhere).

Letting the innovation error covariance \( E = H_S \Sigma H_S^T + R_S \) and the Kalman gain \( K = \Sigma H_S^T E^{-1} \), Eq. (4) can be viewed as a classic “Kalman update” in data assimilation parlance:

### FIG. 1. Conceptual framework for the optimal proxy network design. Tier 1 and tier 2 represent the existing and the unrealized proxy networks. The tier 1 network is assimilated before the optimization procedure (blue arrow) via a proxy system model and the Kalman update equation applied to the climate fields. The greedy optimization (within the blue dashed frame) is then performed over potential sites \( S \) in the tier 2 network.
\[ \mu = \mu + K[y_S - \mathcal{H}_S(\mu)] \quad \text{and} \]
\[ \Sigma = (I - KH_F)\Sigma. \quad \text{(6)} \]

Equation (6) shows that the state error covariance matrix \( \Sigma \) is actually independent of the measurements at sites \( S \). Only \( R_S, \Sigma \), and the mapping operator are necessary to compute \( \Sigma \). Note that \( R_S \) is given by the quality of the observations and \( \Sigma \) is our initial guess on the uncertainty in \( x \).

In this work, we are only interested in some features of \( x \) that we call \( \mathbf{J} \). Let \( \Delta \sigma_f(S) \) denote the change in the variance of \( \mathbf{J} \) due to information at \( S \). Then, using the first-order Taylor expansion of \( \mathbf{J} \), Ancell and Hakim (2007) showed that the assimilation of new observations \( y_S \) will change the covariance of \( \mathbf{J} \) according to

\[ \Delta \sigma_f(S) = \left[ \frac{\partial \mathbf{J}}{\partial \mathbf{x}} \right]^T KH_S \Sigma \left[ \frac{\partial \mathbf{J}}{\partial \mathbf{x}} \right]. \quad \text{(7)} \]

Note that this equation also applies to the generalized case where \( \mathbf{J} \) is a vector, except that \( \Delta \sigma_f(S) \) is a covariance matrix (N. Hryniw and G. J. Hakim 2015, unpublished manuscript). Since the posterior error covariance of \( x \) is independent of \( \mathbf{J} \), \( \Delta \sigma_f(S) \) can be computed as

\[ \Delta \sigma_f(S) = \left[ \frac{\partial \mathbf{J}}{\partial \mathbf{x}} \right]^T KH_S \Sigma \left[ \frac{\partial \mathbf{J}}{\partial \mathbf{x}} \right]. \quad \text{(7)} \]

To estimate the covariance terms required to optimize Eq. (7), we choose an ensemble-based description of the stochastic variables (ensemble sensitivity analysis; Ancell and Hakim 2007; Hakim and Torn 2008) that yields interesting theoretical properties that are very relevant in this context. Submodularity implies that for all subsets \( A \subseteq B \subseteq V \) and element \( s \in V \), the variance reduction function satisfies the following properties:

1) \( \Delta \sigma_f(\emptyset) = 0 \),
2) \( 0 \leq \Delta \sigma_f(A) \leq \Delta \sigma_f(B) \), and
3) \( \Delta \sigma_f(A \cup \{s\}) - \Delta \sigma_f(A) = \Delta \sigma_f(B \cup \{s\}) - \Delta \sigma_f(B) \).

The combinatorial nature of the p-sensor placement problem \( \{O[(q^3 + p^3) \times p_0 C_p]\} \) makes an exhaustive search impractical, such that heuristic methods must be employed, even at the cost of yielding suboptimal solutions. Most variance reduction functions [Eq. (7)] were shown in Das and Kempe (2008) to be submodular, which yields interesting theoretical properties that are very relevant in this context. Submodularity implies that for all subsets \( A \subseteq B \subseteq V \) and element \( s \in V \), the variance reduction function satisfies the following properties:

1) \( \Delta \sigma_f(\emptyset) = 0 \),
2) \( 0 \leq \Delta \sigma_f(A) \leq \Delta \sigma_f(B) \), and
3) \( \Delta \sigma_f(A \cup \{s\}) - \Delta \sigma_f(A) = \Delta \sigma_f(B \cup \{s\}) - \Delta \sigma_f(B) \).

The monotonic increase with the cardinal of \( S \) and submodularity of the reduction function justify the use of a greedy algorithm (Nemhauser et al. 1978) to find near-optimal solutions to our optimization problem. The idea behind the greedy algorithm is simple. It sequentially adds the best sensor location to an existing placement \( S \), starting with the empty set. It has specifically been shown (Nemhauser et al. 1978) that greedy algorithms produce near-optimal solutions when the objective function is nondecreasing and submodular. Near-optimality, in this context, can further be quantified in the sense that the set \( S_G \) resulting from the greedy algorithm is bounded from below by \( (1 - e^{-\lambda}) \Delta \sigma_f(S^e) \), where \( S^e \) is the optimal p-sensor placement and \( e \) is the exponential constant. The total computational cost has thus been replaced by linear growth with the size of the proxy network. With these greedy algorithms available, the exploration of stochastic variability can be carried out in an efficient and comprehensive fashion. A pseudocode of the optimization scheme is shown in Algorithm 1.

**Algorithm 1: Optimal sensor placement—greedy**

Input: \( \Sigma, \mathbf{R}, \mathbf{H}_S, V, p \)
Output: \( S \subset V \) such that \( |S| = p \)

1) \( S \leftarrow \emptyset \)
for \( k \) from 1 to \( p \) do
2) Find \( S^k = \arg\max_{s \in V} \Delta \sigma_f(S \cup \{s\}) \)
3) \( S \leftarrow S \cup \{S^k\} \)

**c. Ensemble Kalman filter**

To estimate the covariance terms required to optimize Eq. (7), we choose an ensemble-based description of the stochastic variables (ensemble sensitivity analysis; Ancell and Hakim 2007; Hakim and Torn 2008) that solves the optimal proxy network problem via an ensemble Kalman filter (EnKF). The approach is similar to the optimal monitoring network design discussed in Mauger et al. (2013).

While the observed values play a crucial role in updating the state estimate in data assimilation, in our case, sensors have to be selected solely based on the sensor characteristics, locations, and their predicted effect on the forecast metric. Initially, \( n \) (independent) samples of \( x \) are used to represent the distribution of the random field, which are assembled as the columns of matrix \( \mathbf{X} \) of size \( q \times n \). The corresponding perturbation matrix \( \delta \mathbf{X} \) is obtained by removing the ensemble mean from each row of \( \mathbf{X} \). We then compute the response function \( \mathbf{J} \) for every realization of \( x \) to get the ensemble forecast \( \mathbf{J} \) of size \( q_0 \times n \) from which the ensemble mean is removed in \( \delta \mathbf{J} \). Then the background error covariance \( \Sigma \) of Eq. (6) can be estimated by the sample covariance:
\[
\Sigma = \frac{1}{n-1} \delta X \delta X^T. \tag{8}
\]

The ensemble covariance estimate of the adjoint sensitivity of \( J \) with respect to changes in the initial state was shown in \( \text{Ancell and Hakim (2007)} \) to be
\[
\frac{\partial J}{\partial x}^T \Sigma = \frac{1}{n-1} \delta J \delta X^T. \tag{9}
\]

Replacing in Eq. (7) the covariance and adjoint sensitivity term with their Monte Carlo (MC) estimates, Eqs. (8) and (9) (i.e., distributions are approximated by ensembles of random samples), we can rewrite the change in the variance of \( J \) due to observations at sites \( S \) [Eq. (7)] as
\[
\Delta \sigma_j(S) = \delta J (H_s \delta X)^T E^{-1} (H_s \delta X) \delta J^T, \tag{10}
\]
where the innovation error covariance is estimated as \( E = H_s \delta X (H_s \delta X)^T + R_s \), and is scalar for a single observation. Combining the greedy optimization approach, which sequentially finds the next best observation location, with the EnKF, which propagates the uncertainty reduction once a site is chosen via the Kalman covariance update, yields an efficient network design tool. Once a site \( s \in V \) is found after one iteration of algorithm 1, the posterior error covariance is computed with Eq. (6) so as to remove the variance explained by observations at \( s \) (\( \text{Mauger et al. (2013)} \)) and is then used as a prior for the next iteration. This differs from the OI approach used in \( \text{Evans et al. (1998b)} \), where stationary covariance is assumed, and hence does not benefit from data-driven updates. In the present study, we found that updating the error covariance significantly affects the resulting error amplitudes (decreasing as more stations were chosen), but not the spatial patterns. This decrease in error covariance amplitude seems to generally lead to a reduced number of chosen replicates when compared to the stationary covariance case. That is, this method is able to extract more information from similarly noisy observations.

Noting that data assimilation is performed for a single site at a time (greedy approach), and to avoid explicit computations with large covariance matrices, we use an equivalent square root formulation (\( \text{Mauger et al. (2013)} \)), which directly updates the perturbation matrix \( \delta X \) rather than its error covariance, as follows:
\[
\delta X_{(k)} = \delta X_{(k-1)} - \beta K [H_s \delta X_{(k-1)}], \tag{11}
\]
where \( k \) denotes the greedy iteration index, \( H_s \) maps \( x \) to observations \( y \) at \( s \), \( r \) is the paleoclimate observational noise variance at site \( s \) connected to the signal-to-noise ratio (SNR) by the relationship \( \text{SNR} = \left( r^{-1} \text{var}(H_s \delta X) \right)^{1/2} \), \( K \) is the Kalman gain expressed as \( K = \delta X (H_s \delta X)^T E^{-1} \), and \( \beta = [1 + (rE)^{1/2}]^{-1} \) results from the square root description of the problem. It reduces the field update according to the proportion of variance at site \( s \) to measurement uncertainty.

Additionally, and to account for the inherent uncertainty that results from approximating the distribution of \( x \) by a climate dataset \( X \), we take a bootstrapping (Monte Carlo) approach, resampling from \( X \) and performing the analysis over \( m \) subsamples \( X_m \) of \( X \) and corresponding response \( J_m \), where \( i \in \{1, 2, \ldots, m\} \) (typically \( m = 1000 \)). Denoting by \( E[\Delta \sigma_j(S)] \), the expected variance reduction for a sensor placement \( S \) over all realizations \( X_m \) of \( X \), then
\[
E[\Delta \sigma_j(S)] = \frac{1}{m} \sum_{i=1}^{m} \Delta \sigma_j(S). \tag{12}
\]

We may now formulate our \( p \)-sensor placement optimization problem as the set \( S \subseteq V \) that maximizes the expected variance reduction over all state matrix realizations \( X \), weighted by their probability of occurrence (here, all equally probable, which assumes a uniform prior):
\[
S^* = \text{argmax}_{S \subseteq V, |S| = p} E[\Delta \sigma_j(S)]. \tag{13}
\]

The OSP formulation described in this article thus differs from that in \( \text{Mauger et al. (2013)} \) because the optimization is performed over the stochastic component, namely the \( X \). Therefore a single sensor placement is obtained and is guaranteed to be near-optimal as explained in section 2b.

Algorithm 1 can thus be modified as follows:

\text{Algorithm 2: Optimal sensor placement—greedy + Kalman filter}
\text{Input:} \( X^{(0)} \rightarrow J^{(0)} \), \( R \), \( V \), \( p \)
\text{Output:} \( S \subset V \) such that \( |S| = p \)
1) \( S \leftarrow \emptyset \)
2) for \( k \) from 1 to \( p \) do
3) \( S \leftarrow S \cup \{s^*\} \)
4) for \( i \) from 1 to \( m \) do
5) \( \text{Update } X_i^{(k)} \) from Eq. (11) and derived \( J^{(k)} \)

3. Application to optimal coral sampling

We now apply this methodology to the placement of coral paleo-observing sites. Using a perfect model approach (e.g., temperature and salinity output fields from GCM simulations), we can quantify the value of coral \( \delta^{18} \text{O} \) record networks in characterizing the uncertainty.
in the climate fields with the sole knowledge of paleoclimate observational error. This is achieved by relying on a PSM, which we describe next.

\[ \delta^{18}O_{m} = \delta^{18}O_{c} + e_{j}, \]  
\[ \delta^{18}O = a_1 \text{SST} + a_2 \text{SSS}. \]  

The model explains large-scale ENSO-related variations in SST and SSS as sampled by a sparse coral $\delta^{18}$O observing network, but underestimates twentieth-century trends in observed coral $\delta^{18}$O. Suppose the measured oxygen isotopic composition of coral aragonite is written

where the subscripts $m$ stand for measured $\delta^{18}$O and $c$ for coral $\delta^{18}$O (the true composition). Also, $j \in \{1, 2, \ldots, p_{0}\}$ is a spatial location index, and $e \sim N(0, R)$ is a measure of nonclimatic influences, whose covariance matrix, of size $p_{0} \times p_{0}$, is assumed diagonal (this assumption can be relaxed if a more detailed structure is suspected). Although the coral PSM was primarily validated for the Indo-Pacific region, we expand its use to a global set of potential sites for the reconstruction of tropical SST field anomaly. We apply the $\delta^{18}$O–SSS relationship for the tropical Atlantic (0.15) and South Atlantic (0.51) regions (LeGrande and Schmidt 2006, their Table 1) to the global potential observing network studied by Evans et al. (1998b). Although the regression of $\delta^{18}$O on SSS in the Atlantic is less skillful than in the Indo-Pacific region, we found our results were not sensitive within reported uncertainty in the regression coefficients (results not shown). For this study, the SNR value was set to 0.5 at all sites, which can be related to the paleoclimate data noise variance $r$ in Eq. (11). While this value is chosen arbitrarily (it probably is a function of, but not limited to, space, time, data type, target variable, and frequency of variation) and lies at the upper range of the spectrum, it is not unrealistic (Smerdon 2012; Wang et al. 2014).

\[ \delta^{18}O_{m} = \delta^{18}O_{c} + e_{j}, \]  
\[ \delta^{18}O = a_1 \text{SST} + a_2 \text{SSS}. \]  

The existing network of subannually resolved corals (mostly in the Indo-Pacific region) documents large interannual and interdecadal variability in sea surface conditions (e.g., Evans et al. 2000; Gagan et al. 2000; Cobb et al. 2008; Lough 2010; Emile-Geay and Eshleman 2013). The following questions motivate the application of the mathematical framework described above:

1) Is the current coral network optimal?
2) Where should the next samples be gathered?
3) Would it be more advantageous to sample at new locations or replicate paleoclimate observations at existing sites?
4) How many paleoclimate observations are needed to yield useful information about a climate field?
5) Can SSS be constrained from corals?
6) When should longer records be chosen over shorter ones?

The model explains large-scale ENSO-related variations in SST and SSS as sampled by a sparse coral $\delta^{18}$O observing network, but underestimates twentieth-century trends in observed coral $\delta^{18}$O. Suppose the measured oxygen isotopic composition of coral aragonite is written

where the subscripts $m$ stand for measured $\delta^{18}$O and $c$ for coral $\delta^{18}$O (the true composition). Also, $j \in \{1, 2, \ldots, p_{0}\}$ is a spatial location index, and $e \sim N(0, R)$ is a measure of nonclimatic influences, whose covariance matrix, of size $p_{0} \times p_{0}$, is assumed diagonal (this assumption can be relaxed if a more detailed structure is suspected). Although the coral PSM was primarily validated for the Indo-Pacific region, we expand its use to a global set of potential sites for the reconstruction of tropical SST field anomaly. We apply the $\delta^{18}$O–SSS relationship for the tropical Atlantic (0.15) and South Atlantic (0.51) regions (LeGrande and Schmidt 2006, their Table 1) to the global potential observing network studied by Evans et al. (1998b). Although the regression of $\delta^{18}$O on SSS in the Atlantic is less skillful than in the Indo-Pacific region, we found our results were not sensitive within reported uncertainty in the regression coefficients (results not shown). For this study, the SNR value was set to 0.5 at all sites, which can be related to the paleoclimate data noise variance $r$ in Eq. (11). While this value is chosen arbitrarily (it probably is a function of, but not limited to, space, time, data type, target variable, and frequency of variation) and lies at the upper range of the spectrum, it is not unrealistic (Smerdon 2012; Wang et al. 2014).

The model explains large-scale ENSO-related variations in SST and SSS as sampled by a sparse coral $\delta^{18}$O observing network, but underestimates twentieth-century trends in observed coral $\delta^{18}$O. Suppose the measured oxygen isotopic composition of coral aragonite is written

where the subscripts $m$ stand for measured $\delta^{18}$O and $c$ for coral $\delta^{18}$O (the true composition). Also, $j \in \{1, 2, \ldots, p_{0}\}$ is a spatial location index, and $e \sim N(0, R)$ is a measure of nonclimatic influences, whose covariance matrix, of size $p_{0} \times p_{0}$, is assumed diagonal (this assumption can be relaxed if a more detailed structure is suspected). Although the coral PSM was primarily validated for the Indo-Pacific region, we expand its use to a global set of potential sites for the reconstruction of tropical SST field anomaly. We apply the $\delta^{18}$O–SSS relationship for the tropical Atlantic (0.15) and South Atlantic (0.51) regions (LeGrande and Schmidt 2006, their Table 1) to the global potential observing network studied by Evans et al. (1998b). Although the regression of $\delta^{18}$O on SSS in the Atlantic is less skillful than in the Indo-Pacific region, we found our results were not sensitive within reported uncertainty in the regression coefficients (results not shown). For this study, the SNR value was set to 0.5 at all sites, which can be related to the paleoclimate data noise variance $r$ in Eq. (11). While this value is chosen arbitrarily (it probably is a function of, but not limited to, space, time, data type, target variable, and frequency of variation) and lies at the upper range of the spectrum, it is not unrealistic (Smerdon 2012; Wang et al. 2014).

\[ \delta^{18}O_{m} = \delta^{18}O_{c} + e_{j}, \]  
\[ \delta^{18}O = a_1 \text{SST} + a_2 \text{SSS}. \]  

The model explains large-scale ENSO-related variations in SST and SSS as sampled by a sparse coral $\delta^{18}$O observing network, but underestimates twentieth-century trends in observed coral $\delta^{18}$O. Suppose the measured oxygen isotopic composition of coral aragonite is written

where the subscripts $m$ stand for measured $\delta^{18}$O and $c$ for coral $\delta^{18}$O (the true composition). Also, $j \in \{1, 2, \ldots, p_{0}\}$ is a spatial location index, and $e \sim N(0, R)$ is a measure of nonclimatic influences, whose covariance matrix, of size $p_{0} \times p_{0}$, is assumed diagonal (this assumption can be relaxed if a more detailed structure is suspected). Although the coral PSM was primarily validated for the Indo-Pacific region, we expand its use to a global set of potential sites for the reconstruction of tropical SST field anomaly. We apply the $\delta^{18}$O–SSS relationship for the tropical Atlantic (0.15) and South Atlantic (0.51) regions (LeGrande and Schmidt 2006, their Table 1) to the global potential observing network studied by Evans et al. (1998b). Although the regression of $\delta^{18}$O on SSS in the Atlantic is less skillful than in the Indo-Pacific region, we found our results were not sensitive within reported uncertainty in the regression coefficients (results not shown). For this study, the SNR value was set to 0.5 at all sites, which can be related to the paleoclimate data noise variance $r$ in Eq. (11). While this value is chosen arbitrarily (it probably is a function of, but not limited to, space, time, data type, target variable, and frequency of variation) and lies at the upper range of the spectrum, it is not unrealistic (Smerdon 2012; Wang et al. 2014).
where \( q \) is the number of field grid points and \( n \) is the number of time realizations. For a single new proxy measurement at location \( s \), \( \mathcal{H} = \mathbf{H}_s \) is a \( 1 \times 2q \) row vector, and in the specific case of \( \delta^{18} \text{O} \) following Eq. (14), it may be written as

\[
\mathbf{H}_s = \begin{pmatrix}
0 & \cdots & 0 & a_1 & 0 & \cdots & 0 & a_2 & 0 & \cdots & 0
\end{pmatrix},
\]

where the \( a_1 \) coefficient is located at index \( s \) and \( a_2 \) at index \( q + s \).

To produce the results that follow, some design choices were required: we denote by \( \mathbf{X}_i \), \( i \in \{1, 2, \ldots, m\} \) for \( m = 1000 \), the subsamples from the last 155 years of the past 1000-yr CCSM4 simulation of SST and SSS fields (i.e., the calibration period is 1851–2005). Since we are mainly interested in reconstructing past SSTs, we choose \( J \) to be the SST field; therefore \( \mathbf{J} \in \mathbb{R}^q \). For vector form \( \mathbf{J} \), we choose to maximize the trace of \( \Delta q_j(S) \) in Eq. (13). The impact of the choice of \( \mathbf{J} \) (\( \mathbf{J} = \{\text{SSS}\} \) or \( \mathbf{J} = \{\text{SST, SSS}\} \) on the OSP is investigated in section 4e. Covariance localization (Gaspari and Cohn 1999) is applied with a correlation length of 6000 km (e.g., covariance is null between points that are more than 12,000 km apart). The pseudoproxy experiments of Smelrod (2012) suggest that global patterns of covariance between paleodata and historical data used for calibrating the paleodata should be tested carefully. Hence, this value was determined, as done in Steiger et al. (2014), by finding the minimum in mean error variance that provides a smooth field. Also, \( V \) is the coral network locations of Table 1, so \(|V| = p_0 = 53 \) sites. In one scenario, the optimization is run such that chosen stations are removed from the list of possible sites after each iteration. In other words, we remove \( \{s^*\} \) from the set \( V \) at line 3 of algorithm 2 and the sensor network size cannot exceed \( p_0 \). In another scenario, \( V \) remains the same during the entire analysis, allowing for sites to be chosen multiple times. The latter scenario introduces possible replicates and we can ask whether it is preferable to sample new sites or multiple replicates at a single location.

d. A two-tiered coral network

There are a few examples of observing networks for specialized phenomena. For the Prediction and Research Moored Array in the Tropical Atlantic (PIRATA) project (Bourlès et al. 2008), the Autonomous Temperature Line Acquisition System (ATLAS) network is localized in order to resolve the two main modes of tropical Atlantic climate variability. Similarly, the Tropical Ocean Global Atmosphere (TOGA) program (McPhaden et al. 2010), and its Tropical Atmosphere Ocean (TAO) network is a major component of the ENSO observing system and is located on a more or less regular array. In all cases, there are limits as to where the sensors can be placed. In our case, the pseudocoral locations were grouped into two tiers (see Table 1 and Fig. 2).

- Tier 1 are locations that approximate the position of publicly available coral \( \delta^{18} \text{O} \) observations (http://www.ncdc.noaa.gov/paleo/corals.html) with a resolution better than 2 months. This is the network used by Emile-Geay and Eshleman (2013).
- Tier 2 are locations where coral samples might potentially be collected. We chose to work with the global potential observing network studied by Evans et al. (1998b), which includes Moorea and the Gambier Islands (French Polynesia), Suwarrow (Cook Islands), Hawaii (Big Island), Andaman Island (eastern India), Lakshadweep (western India), Galapagos (Ecuador), Christmas Islands (Australia in the Indian Ocean), Easter Island (Polynesia in the southeastern Pacific Ocean), the Bahamas and Barbados (Caribbean Sea), Cape Verde Island, St. Helena and Sao Tome (Atlantic Ocean) and American Samoa (South Pacific Ocean) (Evans et al. 1998a).

This network is not exhaustive and while these locations contain corals, those are not necessarily conducive to multicentury-long coral-based reconstructions, but it suffices for a proof of concept. In test cases, the use of an updated database (e.g., ReefGIS; http://reefisreefbase.org) yielded similar results at the granularity imposed by realistic SST/SSS datasets.

e. Network assessment

To evaluate the various sensor networks obtained in the following scenarios, we reconstruct the GCM-simulated SST or SSS fields over the 1700–1850 time period using the pseudoproxies described in section 3b at the candidate sites. Reconstructions are performed via assimilation of the optimal pseudocoral observations, as described in Steiger et al. (2014). At each location, we consider 100 noise realizations to validate our uncertainty quantification scheme (i.e., does the actual coverage rate match the nominal coverage rate? This may be seen as the gray band in Fig. 3a,b, bottom panels).

Next, we quantify reconstruction skill via the point-wise coefficient of efficiency (CE; Cook et al. 1994) for
Table 1. Coral database. Geographic coordinates have been rounded up. (Res is the time resolution in months, and N/A denotes not available.)

<table>
<thead>
<tr>
<th>Record No.</th>
<th>Site</th>
<th>Type</th>
<th>Res</th>
<th>Lat</th>
<th>Lon</th>
<th>Range</th>
<th>Tier</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Abraham Reef</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>22°S</td>
<td>152.5°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>2</td>
<td>Abrolhos</td>
<td>δ¹⁸O</td>
<td>1</td>
<td>28°S</td>
<td>114°E</td>
<td>1794–1994</td>
<td>1</td>
<td>Kuhnert et al. (1999)</td>
</tr>
<tr>
<td>3</td>
<td>Adele</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>15.5°S</td>
<td>123°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>4</td>
<td>Admiralty Island</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>2.6°S</td>
<td>147.5°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>5</td>
<td>Amedee Island</td>
<td>δ¹⁸O</td>
<td>3</td>
<td>22°S</td>
<td>166°E</td>
<td>1660–1993</td>
<td>1</td>
<td>Quinn et al. (1998)</td>
</tr>
<tr>
<td>6</td>
<td>Andaman Island</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>10°N</td>
<td>92.5°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>7</td>
<td>Ascension Island</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>7°S</td>
<td>14.5°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>8</td>
<td>Aqaba</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>29.5°N</td>
<td>35°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>9</td>
<td>Bahamas</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>26.5°N</td>
<td>77.5°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>10</td>
<td>Barbados</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>13°N</td>
<td>59.5°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>11</td>
<td>Belize</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>18°N</td>
<td>87.5°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>12</td>
<td>Bermuda</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>32.5°N</td>
<td>64.5°W</td>
<td>N/A</td>
<td>2</td>
<td>Goodkin et al. (2008)</td>
</tr>
<tr>
<td>13</td>
<td>Barrow Reef</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>20.5°S</td>
<td>115.5°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>14</td>
<td>Bramble Cay</td>
<td>δ¹⁸O</td>
<td>2</td>
<td>9°S</td>
<td>144°E</td>
<td>1774–1993</td>
<td>1</td>
<td>Ault et al. (2009)</td>
</tr>
<tr>
<td>17</td>
<td>Chiriqui</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>8°N</td>
<td>82.5°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>18</td>
<td>Christmas Island</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>10.5°S</td>
<td>105.5°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>19</td>
<td>Cape Verde Island</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>16°N</td>
<td>26°W</td>
<td>N/A</td>
<td>2</td>
<td>Moses et al. (2006a)</td>
</tr>
<tr>
<td>20</td>
<td>Cocos Island</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>5.5°N</td>
<td>87.5°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>21</td>
<td>Diego Garcia</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>7.5°S</td>
<td>72.5°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>22</td>
<td>Easter Island</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>27°S</td>
<td>109.5°W</td>
<td>N/A</td>
<td>2</td>
<td>Mucciarone and Dunbar (2003)</td>
</tr>
<tr>
<td>23</td>
<td>Fehldhu</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>3.5°N</td>
<td>73.5°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>24</td>
<td>Fiji</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>18°S</td>
<td>178.5°E</td>
<td>N/A</td>
<td>2</td>
<td>Linsley et al. (2004)</td>
</tr>
<tr>
<td>25</td>
<td>Funafuti</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>8.5°S</td>
<td>179°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>26</td>
<td>Galapagos</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>1°S</td>
<td>90.5°W</td>
<td>N/A</td>
<td>2</td>
<td>Wellington et al. (1996)</td>
</tr>
<tr>
<td>27</td>
<td>Gambier</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>23°S</td>
<td>134.5°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>28</td>
<td>Gardner</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>25°N</td>
<td>168°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>29</td>
<td>Guam</td>
<td>δ¹⁸O</td>
<td>1</td>
<td>13°N</td>
<td>145°E</td>
<td>1790–2000</td>
<td>1</td>
<td>Asami et al. (2005)</td>
</tr>
<tr>
<td>30</td>
<td>Guaymas</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>28°N</td>
<td>111°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>31</td>
<td>Hawaii</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>21°N</td>
<td>158°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>33</td>
<td>Ishigaki</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>24.5°N</td>
<td>124.5°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>34</td>
<td>Jamaica</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>18.5°N</td>
<td>77.5°W</td>
<td>N/A</td>
<td>2</td>
<td>Haase-Schramm et al. (2003)</td>
</tr>
<tr>
<td>35</td>
<td>Jarvis</td>
<td>δ¹⁸O</td>
<td>3</td>
<td>0°</td>
<td>160°W</td>
<td>1850–1999</td>
<td>1</td>
<td>Ault et al. (2009)</td>
</tr>
<tr>
<td>36</td>
<td>Juan de Nova</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>17°S</td>
<td>42.5°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>37</td>
<td>Kanton</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>3°S</td>
<td>172°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>38</td>
<td>Kapingamarangi</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>0.5°N</td>
<td>154.5°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>39</td>
<td>Kirittimati</td>
<td>δ¹⁸O</td>
<td>1</td>
<td>2°N</td>
<td>157°W</td>
<td>1938–93</td>
<td>1</td>
<td>Evans et al. (1998a)</td>
</tr>
<tr>
<td>40</td>
<td>Kish</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>26.5°N</td>
<td>54°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>41</td>
<td>Kwajalein</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>9.5°N</td>
<td>167.5°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>42</td>
<td>Lakshadweep</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>11°N</td>
<td>73.5°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>44</td>
<td>Laing</td>
<td>δ¹⁸O</td>
<td>3</td>
<td>4°S</td>
<td>145°E</td>
<td>1884–1993</td>
<td>1</td>
<td>Tudhope et al. (2001)</td>
</tr>
<tr>
<td>45</td>
<td>Lombok Strait</td>
<td>δ¹⁸O</td>
<td>1</td>
<td>8°S</td>
<td>116°E</td>
<td>1782–1990</td>
<td>1</td>
<td>Charles et al. (2003)</td>
</tr>
<tr>
<td>46</td>
<td>Madang</td>
<td>δ¹⁸O</td>
<td>3</td>
<td>5°S</td>
<td>146°E</td>
<td>1922–91</td>
<td>1</td>
<td>Tudhope et al. (1995)</td>
</tr>
<tr>
<td>47</td>
<td>Madang</td>
<td>δ¹⁸O</td>
<td>3</td>
<td>5°S</td>
<td>146°E</td>
<td>1880–1993</td>
<td>1</td>
<td>Tudhope et al. (2001)</td>
</tr>
<tr>
<td>49</td>
<td>Maiana Atoll</td>
<td>δ¹⁸O</td>
<td>2</td>
<td>1°N</td>
<td>173°E</td>
<td>1840–1994</td>
<td>1</td>
<td>Urban et al. (2000)</td>
</tr>
<tr>
<td>50</td>
<td>Malo Channel</td>
<td>δ¹⁸O</td>
<td>1</td>
<td>16°S</td>
<td>167°E</td>
<td>1928–92</td>
<td>1</td>
<td>Kilbourne et al. (2004)</td>
</tr>
<tr>
<td>51</td>
<td>Manihiki</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>10.5°S</td>
<td>161°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>52</td>
<td>Maza</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>9°S</td>
<td>143°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>53</td>
<td>Mentawai</td>
<td>δ¹⁸O</td>
<td>1</td>
<td>0°</td>
<td>99°E</td>
<td>1858–1997</td>
<td>1</td>
<td>Abram et al. (2008)</td>
</tr>
<tr>
<td>54</td>
<td>Midway</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>28.5°N</td>
<td>177.5°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>55</td>
<td>Moorea</td>
<td>δ¹⁸O</td>
<td>2</td>
<td>18°S</td>
<td>150°W</td>
<td>1852–1990</td>
<td>1</td>
<td>Boiseau et al. (1998)</td>
</tr>
<tr>
<td>56</td>
<td>Nauru</td>
<td>δ¹⁸O</td>
<td>1</td>
<td>1°S</td>
<td>166°E</td>
<td>1897–1995</td>
<td>1</td>
<td>Guilderson and Schrag (1999)</td>
</tr>
<tr>
<td>57</td>
<td>Nine Mile Reef</td>
<td>δ¹⁸O</td>
<td>N/A</td>
<td>27.5°S</td>
<td>32.5°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
</tbody>
</table>
the 1700–1850 validation period. The CE measures the fraction of common variance between the actual and the reconstructed fields (at each grid point) over the validation period, using the validation period mean-field value as a benchmark. The CE skill score ranges from $-\infty$ to 1, but our maps have cutoff values of $-1$ to 1, where 1 indicates that the reconstructed time series perfectly matches the true sequence, while a negative score denotes that the squared error between the two series exceeds the variance of the true sequence. In some cases, we also juxtapose the tropical average anomaly reconstructions with the true series and look at the ensemble coverage skill. The tropical average is taken over the 35°S–35°N latitude range. To convey the information gain/loss between two networks, we show differences between CE maps ($\Delta$CE), which range from $-2$ to 2.

### Table 1. (Continued)

<table>
<thead>
<tr>
<th>Record No.</th>
<th>Site</th>
<th>Type</th>
<th>Res</th>
<th>Lat</th>
<th>Lon</th>
<th>Range</th>
<th>Tier</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>59</td>
<td>Okinawa</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>26.5°N</td>
<td>128°E</td>
<td>N/A</td>
<td>2</td>
<td>Rosenfeld et al. (2006)</td>
</tr>
<tr>
<td>60</td>
<td>Palau</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>3°N</td>
<td>132°E</td>
<td>N/A</td>
<td>2</td>
<td>Wu and Grotti (2010)</td>
</tr>
<tr>
<td>61</td>
<td>Palmerston</td>
<td>$\delta^{18}$O</td>
<td>3</td>
<td>19°S</td>
<td>169°W</td>
<td>1833–1999</td>
<td>1</td>
<td>Ault et al. (2009)</td>
</tr>
<tr>
<td>62</td>
<td>Palmyra Island</td>
<td>$\delta^{18}$O</td>
<td>1</td>
<td>6°N</td>
<td>162°W</td>
<td>1146–1998</td>
<td>1</td>
<td>Cobb et al. (2003)</td>
</tr>
<tr>
<td>63</td>
<td>Pedra de Lume</td>
<td>$\delta^{18}$O</td>
<td>1</td>
<td>17°N</td>
<td>23°W</td>
<td>1928–2002</td>
<td>1</td>
<td>Moses et al. (2006b)</td>
</tr>
<tr>
<td>64</td>
<td>Pirotan</td>
<td>$\delta^{18}$O</td>
<td>2</td>
<td>23°N</td>
<td>70°E</td>
<td>1948–1989</td>
<td>1</td>
<td>Chakraborty and Ramesh (1998)</td>
</tr>
<tr>
<td>65</td>
<td>Penju</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>4°N</td>
<td>96°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>66</td>
<td>Philippines</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>13.5°N</td>
<td>123.5°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>67</td>
<td>Puerto Rico</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>18°N</td>
<td>66°W</td>
<td>N/A</td>
<td>2</td>
<td>Winter et al. (2000)</td>
</tr>
<tr>
<td>68</td>
<td>Rabaul</td>
<td>$\delta^{18}$O</td>
<td>1</td>
<td>4°S</td>
<td>152°E</td>
<td>1867–1997</td>
<td>1</td>
<td>Quinn et al. (2006)</td>
</tr>
<tr>
<td>69</td>
<td>Rarotonga(2R)</td>
<td>$\delta^{18}$O</td>
<td>1</td>
<td>21°S</td>
<td>160°E</td>
<td>1726–1996</td>
<td>1</td>
<td>Linsley et al. (2008)</td>
</tr>
<tr>
<td>70</td>
<td>Rarotonga(3R)</td>
<td>$\delta^{18}$O</td>
<td>1</td>
<td>21°S</td>
<td>160°E</td>
<td>1874–2000</td>
<td>1</td>
<td>Linsley et al. (2006)</td>
</tr>
<tr>
<td>71</td>
<td>Rarotonga(99)</td>
<td>$\delta^{18}$O</td>
<td>2</td>
<td>21°S</td>
<td>160°E</td>
<td>1906–99</td>
<td>1</td>
<td>Linsley et al. (2006)</td>
</tr>
<tr>
<td>72</td>
<td>Ras Umm Sidd</td>
<td>$\delta^{18}$O</td>
<td>2</td>
<td>28°N</td>
<td>34°E</td>
<td>1751–1995</td>
<td>1</td>
<td>Felis et al. (2000)</td>
</tr>
<tr>
<td>73</td>
<td>Rocas Reef</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>4°S</td>
<td>33.5°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>74</td>
<td>Samoa</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>14.5°S</td>
<td>171°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>75</td>
<td>Savusavu</td>
<td>$\delta^{18}$O</td>
<td>2</td>
<td>17°S</td>
<td>179°E</td>
<td>1939–2001</td>
<td>1</td>
<td>Bagnato et al. (2004)</td>
</tr>
<tr>
<td>76</td>
<td>Secas</td>
<td>$\delta^{18}$O</td>
<td>1</td>
<td>8°N</td>
<td>82°W</td>
<td>1707–1984</td>
<td>1</td>
<td>Linsley et al. (1994)</td>
</tr>
<tr>
<td>77</td>
<td>Seychelles</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>4.5°S</td>
<td>55.5°E</td>
<td>N/A</td>
<td>2</td>
<td>Levy et al. (2006)</td>
</tr>
<tr>
<td>78</td>
<td>Solomon Islands</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>8°S</td>
<td>157°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>79</td>
<td>Somalia</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>1°S</td>
<td>42°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>80</td>
<td>Spratly Island</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>8°N</td>
<td>113°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>81</td>
<td>St. Helena</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>16°S</td>
<td>6°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>82</td>
<td>Sao Tome</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>0°</td>
<td>6.5°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>83</td>
<td>South Sulawesi</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>5.5°S</td>
<td>123°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>84</td>
<td>Suwarrow</td>
<td>$\delta^{18}$O</td>
<td>3</td>
<td>13°S</td>
<td>163°W</td>
<td>1880–1999</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>85</td>
<td>Tahiti</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>17.5°S</td>
<td>149.5°W</td>
<td>N/A</td>
<td>2</td>
<td>DeLong et al. (2010)</td>
</tr>
<tr>
<td>86</td>
<td>Tanzania</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>7°S</td>
<td>38°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>87</td>
<td>Tarawa</td>
<td>$\delta^{18}$O</td>
<td>1</td>
<td>1°N</td>
<td>172°E</td>
<td>1894–1990</td>
<td>1</td>
<td>Cole et al. (1993)</td>
</tr>
<tr>
<td>88</td>
<td>Thitu Island</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>11°N</td>
<td>114°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>89</td>
<td>Truk</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>7.5°N</td>
<td>152°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>90</td>
<td>Tutia</td>
<td>$\delta^{18}$O</td>
<td>2</td>
<td>8°S</td>
<td>39°E</td>
<td>1847–1999</td>
<td>1</td>
<td>Ault et al. (2009)</td>
</tr>
<tr>
<td>91</td>
<td>Veracruz</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>20°N</td>
<td>92°W</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>92</td>
<td>Yap</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>9.5°N</td>
<td>138°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>93</td>
<td>Yemen</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>14°N</td>
<td>43°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>94</td>
<td>Yongala Reef</td>
<td>$\delta^{18}$O</td>
<td>N/A</td>
<td>19.5°S</td>
<td>147°E</td>
<td>N/A</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>95</td>
<td>Zanzibar</td>
<td>$\delta^{18}$O</td>
<td>1</td>
<td>6°S</td>
<td>39°E</td>
<td>1851–1999</td>
<td>1</td>
<td>Ault et al. (2009)</td>
</tr>
</tbody>
</table>

### 4. Results on synthetic data

We now address the motivating questions posed in section 3a with this virtual network.

#### a. Is the current coral network optimal?

In this experiment, we compare the performances of the pseudocoral $\delta^{18}$O network at existing locales (33 unreplicated tier 1 locations listed in Table 1) with the optimal 33-sensor network. Again, the target is the 1700–1850 SST field. Figures 3a and 3b (top panels) illustrate how the CE skills of the two networks differ spatially. Looking at the areas of positive CE, the OSP is more efficient in the Indonesian and Philippine regions, the northern tropical and central Pacific, and the southern tropical Atlantic. This point is reinforced in the bottom panel of Fig. 3, which displays the pointwise
difference between the CE map for the optimal 33 sensors (Fig. 3b, top) and the CE map for tier 1 (Fig. 3a, top). The OSP skill is seen to exceed the tier 1 skill in the regions cited above. It is, however, lower in the southwestern Pacific region and the southeastern Indian Ocean. [Looking at Fig. 6 (top, red dot and blue curve), the optimal 33 sensors explain nearly 10% more SST variance than tier 1.] This demonstrates the advantage of our design approach, which allows us to quantify the information added by a particular set of sensors. Although the field skill is quite different, the skill is indistinguishable for the reconstructed tropical means. Despite the warm bias (von Storch et al. 2004; Smerdon et al. 2011; Wang et al. 2014), which is due to the calibration period 1851–2005 being warmer than the reconstructed interval [1700, 1850], both networks perform well in reconstructing the tropical average SST and have uncertainty bands of similar width. The two networks differ spatially in that the OSP privileges the northern over the southern tropical Pacific sites, with the exception of the Gambier Islands (French Polynesia). Atlantic locations are also picked more often in the OSP case. Sites around Australia, however, are not being chosen, probably due to their localized spatial impact. This result, however, is model dependent since the covariance structure depends on the target SST and SSS fields (the CCSM4 “past1000” simulation). Different results may be obtained with different target fields.

b. Where should the next samples be gathered?

Given the current network of 33 unreplicated sites (tier 1), where should the next 25 or 50 coral records be gathered? As in the previous experiment, we do not allow the sampling of replicates. Using an SNR of 0.5, Fig. 4 shows how the SST reconstructions improve with respect to the addition of \( p = 0, 25, \) and \( 50 \) records.

Figure 4 (top) shows the pointwise CE skill score for tier 1, while Fig. 4 (middle and bottom) illustrate the incremental increase of the positive CE area with the network size via \( \delta CE \), that is, the CE gain from adding 25 and 50 optimally placed sensors to tier 1 (Fig. 4, middle and bottom). Note that \( \delta CE \) clearly demonstrates the diminishing return property, as the first additional 25 optimal sites have a greater impact on the positive CE area (Fig. 4, middle) than the subsequent 25 (Fig. 4, bottom). The next sites to sample for SST reconstruction are the Gambier Islands (French Polynesia), followed by the southern tropical Atlantic region, the equatorial Indian Ocean, and the northern tropical Pacific sites. However, the mean SST field variance reduction seems to stall as the network size increases (Fig. 6, bottom; blue curve). As we add 65 new sites to tier 1, the variance reduces by less than 20%; hence, we hypothesize that a better strategy is to allow replication to reduce the paleoclimate observational uncertainty.

c. Are replicates more advantageous?

We repeat the previous analysis, this time allowing for a site to be chosen multiple times. Given the current pseudocoral network of 35 sites, which includes two replicates in Rarotonga (Cook Islands), where should we sample the next 50 or 100 corals, allowing for replication? We call this scenario our “control” run for future comparison. Results are shown in Fig. 5.

For up to 100 paleoclimate observations on top of the 35 existing sites, the addition of sensors greatly improves the SST field reconstruction. The combined positive \( \delta CE \) area in Fig. 5 (middle and bottom) showing the information gain from adding the optimal 50 and 100 sensors to tier 1, covers nearly the entire Pacific, Indian, and Atlantic Oceans with the exception of the southern and northern limits of the map. We note a reduction of the mean SST field variance of 35% when adding 100 optimally placed sensors to tier 1 (Fig. 6, bottom; green curve). The resampling rate is highest in the central Pacific region (Palmyra and Kiritimati) as well as in the eastern Pacific Ocean (Galapagos and Clipperton Atoll), followed by the Seychelles in the Indian Ocean. Note also that replicates occur in places where there are few nearby alternatives.
How many paleoclimate observations are needed to yield useful information about a climate field?

OSP results are highly sensitive to paleoclimate observational error, or equivalently, SNR [Eq. (11)]. We illustrate this by experiments which achieve a similar level of reconstruction CE (Fig. 7) while sampling plausible ranges of SNR and numbers of sites/replicates. Figure 7 shows that similar CE patterns and amplitudes can be achieved by 10 sites sampled with SNR $= 5$ (Fig. 7, top), 25 sites sampled with SNR $= 0.5$ (Fig. 7, middle), or 100 sites sampled with SNR $= 0.25$ (Fig. 7, bottom). Focusing on the Niño-3.4 region in Fig. 7, note how the number of replicates scales up as the SNR decreases—growing from 3 to 8 to 30 for SNR = 1, 0.5, and 0.25 respectively. It shows that a few chosen high-quality records can carry more information about the SST field as a whole, and hence represent a better investment of resources, than a much greater number of low-quality records.

FIG. 3. Is the current coral network optimal? Evaluation of the SST field from 1700 to 1850 reconstructed using (a) an existing and (b) our optimal $p$-proxy network for $p = 33$: (top) pointwise CE; (middle) true and reconstructed tropical mean SST field anomaly with the shaded area being their 95% confidence bounds. The grayscale dots in the top panel of (b) represent the paleoclimate observation sites in impact order, black being most impactful and white least impactful. The blue dots in the top panel of (a) show the existing coral network locations. (bottom) Difference map between the optimal 33-sensor network CE map and tier1 CE map.
e. Are corals more informative of temperature or salinity?

So far we have focused on how much temperature information could be gleaned from a sparse network of coral observations. Could such a network also inform us about salinity (SSS)? If so, how would an optimal SSS-characterization coral network differ from an optimal SST-characterization coral network? Is it possible to jointly estimate SST and SSS with the same network?

To address these questions, we repeat the experiment carried out in section 4c with two alternative selections.

Fig. 4. Where should the next samples be gathered? (top) Pointwise CE between the reconstructed and true SST fields from 1700 to 1850 using tier 1; also shown are difference maps \( \delta \text{CE} \) when adding (middle) 25 and (bottom) 50 optimal sensors to tier 1. The dots in grayscale represent the optimal paleoclimate observation sites in impact order, black being most impactful and white least impactful. The blue dots show the existing coral data locations.
for $J$, namely $J = \{\text{SSS}\}$ (Fig. 8) and $J = \{\text{SST, SSS}\}$ (Fig. 9). The coral network that minimized the uncertainty in the SSS field (Fig. 8) seems much more concentrated around the western Pacific warm pool (Philippines and Papua New Guinea) and sparser in the central Pacific region (Fig. 8); this is ostensibly because salinity is only a dominant influence on coral $\delta^{18}O$ in the western Pacific warm pool region, where SST variations are weak compared to precipitation-induced SSS variations. Note also that the SSS field is more difficult to reconstruct accurately with tier 1 and tier 2 networks and requires many more sensors, perhaps not limited to...
FIG. 6. Mean SST field variance reduction as a function of the network size. (top) Variance reduction for the three networks independently and (bottom) variance reduction for tier 1 + OSP. The red circles correspond to the variance reduction before and after the assimilation of the entire tier 1 network at once. The blue curve is for the unreplicated OSP and the green curve is for the OSP allowing replicates.
cultural records. There is, however, little structural difference between the OSP obtained with $J = \{\text{SST}, \text{SSS}\}$ (Fig. 9) and that using $J = \{\text{SST}\}$ only (Fig. 5). This indicates that, with this synthetic network PSM at least, coral $\delta^{18}$O is so dominated by thermal signals that little additional information can be extracted about SSS.

f. How important is the length of the records?

So far we have only considered an extremely idealized, complete paleoclimate observational coverage. In reality, most series are of different lengths, and we would like to quantify how this affects site selection. We assume a time span that is common to all the sites between 1931 and 2005, but a random start date (chosen from a uniform distribution over 1851–1930) is assigned to each location such that some samples go farther back in time than others. A single realization of start dates is used in this experiment. Therefore, some samples may provide at most twice as many annual measurements as others. In Eq. (11), the length of a particular time series specifies the ensemble size that is being updated, so as longer sequences are being assimilated, those affect the update of a larger portion of the ensemble, while shorter ones only impact a subsample thereof. Additionally, short sequences result in noisier paleoclimate observations due to the constant SNR. This setting would be representative of cases where the average length of samples is known in specific areas. The resulting OSP
(Fig. 10, middle), when compared with the OSP obtained with fixed length sequences (Fig. 10, top) is different as it generally exhibits a lesser replication rate. Where the fixed lengths OSP picks 19 distinct sites, the variable lengths OSP choses 25 sites that seem to be more uniformly distributed within the climatically important regions. Yet, not all the chosen pseudocorals have long sequences (Fig. 10, bottom) suggesting that location might be just as important as length.
5. Discussion

The previous section laid down a proof of concept for how the data assimilation framework may be used to design optimal sampling strategies and answer broad scientific questions. Because it uses a linear bivariate forward model instead of assuming a univariate dependence on temperature with 1:1 scaling, this work brings the exercise of Evans et al. (1998b) one step closer to reality. It also approaches optimality of the total resultant network, rather than an iterative next best site to explore, and carries information gained sequentially as sites are being assimilated. Finally, its computational efficiency allows for the testing of multiple scenarios.

Fig. 9. OSP performance using $J = \{SST, SSS\}$. (top) Pointwise CE between the reconstructed and true SST fields from 1700 to 1850 using tier 1 (with replicates); also shown are difference maps $\delta CE$ when adding (middle) 50 and (bottom) 100 optimal sensors to tier 1. The dots in grayscale represent the optimal paleoclimate observation sites in impact order, black being most impactful and white least impactful. The blue dots show the existing coral data locations. The numbers shown within the dots represent the number of replicates at the site.
However, results from this optimization problem are clearly sensitive to choices made to close the problem. In this section, we discuss some of our assumptions and how they may be relaxed to provide more realistic results in operational use. We also consider some possible extensions.

a. Assumptions

1) REALISTIC TARGETS

The CCSM4 simulation that generated the target SST and SSS fields used here is a plausible representation of tropical ocean–atmosphere processes on time scales of interest and produced long (150 yr) calibration and validation periods, necessary to provide reliable estimates of the error covariances. However, its spatial covariance structure differs from the true SST and SSS (e.g., Deser et al. 2012). For a more realistic sampling design experiment, one might use historical observations to represent the true target fields as closely as possible. Those, by contrast, are spatially and temporally incomplete (therefore biased), noisy, and span a shorter time range, which may require some form of regularization (space reduction) to stabilize covariances at the expense of spatial detail (Kaplan et al. 1997). Additionally, the cost functions used in most of our experiments are for (co)variance of the SST and/or SSS fields. We used the entire SST or SSS fields as targets for generality, but the approach is applicable to any derived function of those fields. Thus, our algorithm would work just as well at a finer spatial resolution, if one chose to focus on a regional process or an area index (e.g., Niño-3.4 SST anomaly or SSS averaged over the Caribbean Sea).
2) PSM DEPENDENCY

The optimization results depend on the PSM chosen to model the paleoclimate observations and whether it precisely and accurately represents the response of the sensor to SST and SSS variations. Given the emerging state of the proxy system modeling field, our analysis only considers a single PSM, for which we fixed the parameters to the values proposed by Thompson et al. (2011). However, in reality one needs to account for structural uncertainties as well as parametric uncertainties. Parametric uncertainties could be tackled with sensitivity analyses using different values for the model parameters. Structural uncertainties could be identified with experiments comparing validated performance of a hierarchy of PSMs spanning a broad range of structural complexities.

3) LIST OF CANDIDATE SITES

Because of the steep and anisotropic gradients in SST/SSS, results might also be sensitive to the exact location of paleoclimate observational sites in tier 1 and tier 2. For operational applications and depending on the granularity of interest, one might want to use an updated network of potential sites (e.g., ReefBase; http://reefgis.reefbase.org), in conjunction with an observational target with similarly enhanced spatial and/or temporal resolution.

4) SNR ESTIMATION

In this paper, for simplicity, we used a uniform SNR value of 0.25, 0.5, or 1.0 for all potential sites, but this assumption can be relaxed and different SNR values may be specified at each location. If we were to perform an experiment with spatially varying SNR values, experience shows that, other things being equal, sites with high SNRs would be favored by our algorithm. In the real world, empirical SNR values are challenging to assign, especially given that paleoclimate observational uncertainties at different sites might include both random and systematic errors. In the case where in situ or local calibrations using contemporaneous direct and indirect observations have already been performed at the site (e.g., the existing tier 1 network), the SNR may be directly estimated (e.g., Evans et al. 2000). If that is not the case, then one must make an educated guess given SNR values observed across different samples, such as drawing SNR values from empirical SNR distributions (e.g., Wang et al. 2014).

b. Extensions

1) CHOICE OF COST FUNCTION

The choice of cost function is crucial to the sensor selection process. Because our goal is to predict the information utility of specific proxy measurements before obtaining the data, the cost function evaluates the compactness, spread, and variance of the estimated targets. In this study, we used the total variance reduction for $J$, but we could also use the entropy of $J$ conditioned on observations at sites $S$. However, the entropy reduction, unlike the variance reduction, is neither submodular nor supermodular (Anstreicher et al. 2001) and prevents the use of a greedy algorithm. Evans et al. (1998b) also optimized based on resolution of area-weighted indices such as the Niño-3.4 index, the global temperature, and the warm pool average. Alternatively, Krause et al. (2008b) use a mutual information maximization criteria between sites $S$ and the rest of the space $V \setminus S$ approximated as

$$I(S; V \setminus S) = \frac{\sigma_j^2 - \Delta \sigma_j(S)}{\sigma_j^2 - \Delta \sigma_j(V \setminus S)},$$

which allows them to reduce the error about the estimates over the space not covered by the sensor locations. They were also able to show that the mutual information criteria were submodular, and that for sparse enough observation densities they were approximately nondecreasing, which permits the use of a greedy optimization approach. We leave the investigation of the OSP sensitivity to the choice of the cost function for future work.

2) ANCILLARY OPTIMIZATION CONSTRAINTS

We have assumed that practical considerations (e.g., differences in site access, sample availability, systematic differences in longevity of coral colonies depending on the mean temperature and climatic variability, distribution of suitable species) are uniform across candidate sites. We could allow for more realism by considering a parameterized deployment cost for each potential observing site. Here we considered the potential coral sites from Evans et al. (1998b) according to a priori biological and geological considerations, but clearly those do not encompass all suitable locations. Conversely, not all of those sites may actually host coral species (e.g., *Porites*) and colonies suitable for climate reconstruction, for instance because of recent environmental deterioration, tropical cyclone activity, or simply because suitable materials are not safely or practically accessible. Specifically, corals at tier 2 locations such as Cape Verde, Easter Island, Rocos Reef were shown to be unfit for paleoclimate reconstruction (see Moses et al. 2006a; Mucciarone and Dunbar 2003; Ferreira et al. 2013, respectively). On the other hand, some locations are known to provide particularly relevant coral records for reconstructions (e.g., in Bermuda, Fiji, Puerto Rico, and La Reunion; see respectively Goodkin et al. 2008; Linsley et al. 2004; Winter et al. 2000; Pfeiffer et al. 2004). Therefore, as a step toward realism, expert judgement regarding coral quality at potential sites could
easily be included in our analysis in the form of an optimization constraint. We might do so by assigning a weight \( w_i \in [0, 1] \), where \( i = 1, \ldots, p \), assessing conditions at each site. Sites at which corals previously produced valuable reconstructions could be given higher weights, while sites where corals were reported to be stressed or unfit would be downgraded. The algorithm would favor sites weighted more heavily, and this would have the added advantage of making site selection criteria explicit and transparent. Since results may be sensitive to weighting, efforts should be made to assess the validity of the weighting assignments before making inference on the results.

3) CHRONOLOGICAL UNCERTAINTIES

This work has neglected age uncertainties, which are known to exist in corals and in all paleoclimatic observations, for which time of formation or imprinting is measured (Evans et al. 2013). The effect of chronological uncertainty on observing network definition may be assessed by incorporating ensemble estimation of plausible chronologies of the paleoclimatic data using probabilistic modeling (e.g., Comboul et al. 2014).

4) APPLICATION TO OTHER OBSERVING NETWORKS

Although we focused our results on one proxy system (coral \( \delta^{18}O \)), we note that this framework is generally applicable to any proxy system that is well understood enough to be explicitly modeled via a PSM. An open-source toolbox of such PSMs (Dee et al. 2015) will greatly expand the scope of such sensor placement studies. In our coral example, we used a linear relationship \( \langle H \delta\rangle \) between the climate fields and the network of potential paleoclimatic observations. However, one could use any nonlinear anomaly mapping \( \delta^{18}O = \langle H \delta\rangle - \langle \delta^{18}O \rangle \), where \( \langle H \delta\rangle \) denotes a nonlinear operator and the angle brackets the ensemble mean, in lieu of the linear relation \( \langle H \delta\rangle \). Because the mapping is site specific, both the PSM parameters and structures may vary across a network, permitting optimizations for observing networks comprised of multiple classes of paleoclimatic sensors and observations (Emile-Geay et al. 2013a,b).

In practice, one could apply this methodology to the design of observing networks for multivariate, indirect responders to environmental conditions with observational error of any sort; thereby enabling applications in radioastronomy, fixed buoy networks, Lagrangian ARGO floats, or satellite remote sensing arrays.

6. Conclusions

We have described a general method to solve the optimal paleoclimatic sensor placement problem and illustrated its use for the design of coral \( \delta^{18}O \) sampling networks using simulated climate fields. Although the subject had been investigated earlier by Evans et al. (1998b), here we presented a more generalized and realistic approach to its closure that was permitted by the integration of three novelties: 1) a data assimilation method, consisting of the square root formulation of the EnKF (Mauger et al. 2013), to skillfully measure the (possibly multidimensional) information gain from paleo-observing networks; 2) a regionally defined bivariate PSM relating SST and SSS anomaly fields to coral \( \delta^{18}O \) (Thompson et al. 2011); and 3) an effective optimization scheme, which is a greedy algorithm (Nemhauser et al. 1978), enabled by exploiting the submodular and monotonic properties of the variance reduction criteria. While our optimization results are likely to depend on the GCM, PSM, target network, and assumptions chosen, we demonstrated the general merit of the OSP approach by quantifying the impact of additional paleoclimate observations on the reconstruction of simulated preinstrumental climate fields and by exploring several design strategies. In the idealized situation presented here (perfect climate model, perfect coral system model), we gave the following responses to our motivating questions (section 3a):

1) Is the current coral network optimal?

The existing network overly samples the south west Pacific region. The OSP approach targets more sparsely sampled regions, thereby achieving a more uniform coverage of the reconstruction domain.

2) Where should the next samples be gathered?

Given the existing coral \( \delta^{18}O \) network (i.e., tier 1), our list of potential sites (tier 1 + tier 2), and the goal to maximize knowledge of the tropical SST field at \( 5^\circ \times 5^\circ \) resolutions, the next samples should be sought first in French Polynesia, then the southern tropical Atlantic, followed by the equatorial Indian Ocean and the northern tropical Pacific. By contrast, maximizing knowledge of the tropical SSS field yields heavily to new samples located in the southern tropical Pacific.

3) Are replicates more advantageous?

The choice of optimal sites depends sensitively on the initial assumption regarding the uncertainty in the paleoclimatic observations (sensor precision; Evans et al. 1998b). While high-quality paleoclimate observations lead to mostly unreplicated OSPs (SNR \( \geq 1 \)), replicates prove advantageous when errors are larger, in which case they essentially improve the SNR via the central limit theorem (assuming the observational error is random and not systematic). The replication rate is highest in

Unauthenticated | Downloaded 09/01/22 11:27 PM UTC
the central equatorial Pacific (Palmyra and Kiritimati) and the eastern Pacific (Galapagos and Clipperton Atoll) (Evans et al. 1998b) to infer both the SST field alone and the SST and SSS fields combined. Inferring the salinity field alone results in replicates that are more uniformly distributed within the western Pacific warm pool, besides a few places such as in the Cook Islands and the Seychelles.

4) How many paleoclimate observations are needed to yield useful information about a climate field?

This experiment and the previous one lead to a similar conclusion in that the quality of the paleoclimate observations determines the size of the paleo-observing network.

5) Are corals more informative of temperature or salinity?

We showed that to infer the SSS resulted in a sensor network that was structurally different from the OSP inferring the SST. However the OSP for the combined SST and SSS fields was very close to the one computed with the temperature field, because the contribution to J of δ 18O variance by the scaled SST dynamic range (section 3b) is larger than that for the scaled SSS dynamic range. In combination for the tendency for SSS and SST to covary in the tropical climate, these results support the use of coral δ 18O records to reconstruct SST (see Tierney et al. 2015).

6) How important is record length?

Record lengths played a major role in the OSP design, with variable pseudocoral time series length encouraging more uniformly distributed networks and less replication.

The idealized nature of these experiments was necessary to establish a proof of concept, so these results are merely illustrative of the type of questions one may address via this framework; they should not be taken literally. Some of these results (items 3–6 above) are general enough that they should carry over to real world. However, the first two questions hinge crucially on the network of potential sites, resolution, and the spatial relations simulated by the GCM, which are known to be unrealistic in CCSM4 and other GCMs from phase 5 of CMIP (CMIP5) because of the so-called “cold tongue” bias (Li and Xie 2014), among other things [see Bellenger et al. (2014) for a recent review].

Operationalizing the OSP framework for paleoclimatic applications could be done relatively easily and accommodate a wide range of applications. It would require 1) the use of instrumental data as target climate fields; 2) validated PSM(s) (ideally, evaluating sensitivity to parametric and structural uncertainties); 3) an accurate catalog of potential sampling sites defined using comprehensive and updated paleoclimatic and ecological databases, on an appropriately resolved target field; and 4) estimation of SNR values using existing, validated paleoclimate record calibration studies. In future work, we may apply this framework to real-world optimal coral sampling, and extend it to other classes of paleoclimatic observations.

Acknowledgments. We thank all reviewers and the Editor for constructive reviews that improved the paper. JEG and MC acknowledge funding from NOAA awards NA10OAR4310115 and NA14OAR4310175. GH acknowledges funding from NSF award AGS-1304263 and NOAA award NA14OAR4310176.

REFERENCES


