



HAL
open science

Reconstructing the Ocean State Using Argo Data and a Data-Driven Method

Erwan Oulhen, Nicolas Kolodziejczyk, Pierre Tandeo, Bruno Blanke, Florian Sévellec

► **To cite this version:**

Erwan Oulhen, Nicolas Kolodziejczyk, Pierre Tandeo, Bruno Blanke, Florian Sévellec. Reconstructing the Ocean State Using Argo Data and a Data-Driven Method. *Journal of Atmospheric and Oceanic Technology*, 2024, pp.1-30. 10.1175/JTECH-D-23-0157.1 . hal-04759542

HAL Id: hal-04759542

<https://imt-atlantique.hal.science/hal-04759542v1>

Submitted on 29 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

1 **Reconstructing the Ocean State Using Argo Data and a Data-Driven Method**

2 Erwan Oulhen,^a Nicolas Kolodziejczyk,^a Pierre Tandeo,^{b,c,d} Bruno Blanke,^a Florian
3 Sévellec,^{a,c}

4 ^a *Univ Brest, CNRS, IRD, Ifremer Laboratoire d'Océanographie Physique et Spatiale (LOPS,*
5 *UMR 6523), IUEM, Brest, France*

6 ^b *IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238, Brest, France*

7 ^c *Odyssey, Inria/IMT/CNRS, Rennes, France*

8 ^d *RIKEN Center for Computational Science, Kobe, 650-0047, Japan*

9 *Corresponding author: erwan.oulhen@univ-brest.fr*

10 ABSTRACT: Twenty years of Argo profiles have provided unprecedented insight into the global
11 space and time variability patterns of ocean temperature and salinity, significantly reducing asso-
12 ciated uncertainties. However, such assessments during the pre-Argo period remain a challenge
13 due to the scarcity of observations in many regions. From the Argo period, a set of dominant
14 three-dimensional patterns can be estimated using EOF analysis and used to fill in observational
15 gaps. From the associated principal components, temporal fluctuations can be observed, aiming
16 to build a catalog of possible trajectories of the ocean state. To map pre-Argo observations,
17 EOFs are used in a data assimilation framework using a catalog to feed an analog prediction and
18 provide reanalysis. In this study, this new data-driven interpolation method is called RedAnDA
19 (Reduced-space Analog Data Assimilation) and was tested in the tropical Pacific Ocean. RedAnDA
20 was first validated through an Observing System Simulation Experiment (OSSE) approach, using
21 synthetic observations extracted from a model simulation. It was then applied to a real historical
22 dataset and compared to other available reanalysis products. Overall the reconstructed temperature
23 field showed variability consistent with the OSSE application and other reanalysis products in the
24 real data application. Further improvements are needed to optimally estimate uncertainty, but
25 RedAnDA already combines valuable information about state predictability, observation sampling,
26 and unresolved scale issues.

27 1. Introduction

28 In the ocean, temperature, by setting the density in most of tropical to subpolar regions, plays
29 a fundamental role in influencing dynamical processes such as three-dimensional geostrophic
30 circulation and the mixing of water masses. Temperature has a first order impact on ecosystems
31 (?). The globally averaged Ocean Temperature temporal changes reflect long-term global warming
32 through heat uptake (??), but also climate and atmospheric variability by air-sea interaction (?).
33 Moreover, the ocean exhibits significant variability spanning regional to basin scales and ranging
34 from seconds to decades, which can superimpose and interact, increasing uncertainties in estimating
35 long-term anthropogenic trends or underlying natural variability (???). Notably, phenomena such
36 as the air-sea coupled El Niño-Southern Oscillation (ENSO) have a pronounced impact on surface
37 temperature primarily in the Pacific Ocean, with global impacts on long-term temperature trends
38 (??). Consequently, evaluating the influence of natural oceanic temperature variability represents
39 a crucial challenge. Therefore, precise estimates of global and regional oceanic variability derived
40 from *in situ* measurements remain timely.

41 From the late 1930s to the 1990s, successive advances in *in situ* measurement technology
42 facilitated the expansion of observational sampling of the ocean, particularly along repeated hy-
43 drographic transects in the major ocean basins and sub-basins. However, a notable bias persisted
44 in the distribution of observations, in favor of a better sampling the Northern Hemisphere and
45 coastal regions (?). Furthermore, the majority of research expeditions were conducted during
46 summer months, primarily due to favorable weather conditions for navigation. This lack of *in*
47 *situ* observations has impeded our ability to consistently describe regional patterns and understand
48 underlying oceanic mechanisms (?), thereby contributing to uncertainties in quantifying the rate
49 of ocean heat content change (?????).

50 Since the late 1990s, the Argo global observing system has been deployed, achieving in 2007,
51 its targeted quasi-global coverage of 3x3° T and S profiles over the upper 2000 m of the ocean
52 (??). Over the last two decades, it has become the backbone of the Global Ocean Observing
53 System (GOOS) and is now the primary source of observation for numerous ocean and climate
54 studies (see ?, ?, for a recent review). The Argo system effectively addresses the major sampling
55 biases inherent in ocean observing systems, reducing uncertainties in global oceanic records and

56 enhancing the confidence in assessments of ocean heat content, thermal expansion, and changes in
57 freshwater content (Hansen et al. 2011; ?; ?; ?).

58 One of the principal challenges in oceanography lies in filling the gaps in sparsely distributed
59 observations, especially during the pre-Argo period, i.e., before 2000. Monitoring of the global
60 ocean state is significantly enhanced when observational variables are remapped on a regular grid
61 utilizing statistical methodologies or assimilated into a numerical model. Among the procedures
62 for mapping oceanic data, several methods can be categorized.

63 Firstly, widely used mapping approaches are methods that rely solely on observations and *a priori*
64 statistics, such as Optimal Interpolation (OI, Bretherton et al. 1976). Nevertheless, they exhibit
65 limited skill in accurately estimating error. Among the notable Optimal Interpolation applications,
66 Good et al. (2013) reconstructed past temperature and salinity variability, from 1900 onwards, with
67 a focus on evaluating observational quality. Kaplan et al. (1997) combined data reduction using
68 Empirical Orthogonal Functions and least squares mapping techniques (e.g., optimal smoother,
69 Kalman filter, and OI) to analyze Sea Surface Temperature (SST) anomalies in the Atlantic Ocean
70 for the period 1856-1991. This approach was computationally efficient and yielded consistent
71 results with robust error estimates. A last example is ?, who utilized an ensemble OI approach,
72 incorporating a first guess and *a priori* covariance derived from the 5th phase of the Coupled Model
73 Intercomparison Project (CMIP5), to reconstruct ocean subsurface temperature patterns and the
74 trend from 1940 to 2014.

75 Secondly, data assimilation procedures in numerical models allow for a physically consistent
76 data interpolation (?). These analyses give robust estimates of upper Ocean Heat Content (OHC)
77 and the depth of the tropical mixed layer (?). However, this approach requires significant numerical
78 costs, as it requires the integration of the primitive equations into an Ocean General Circulation
79 Model (OGCM) and are subject to inherent numerical imperfections (?).

80 In recent years, a new class of methods has emerged, taking advantage of the ever-increasing
81 amount of data available from both satellite and *in situ* measurements, and recent advances in
82 applied mathematics. These data-driven methods have been specifically developed for data assimi-
83 lation and climate prediction purposes (e.g., ?????). Leveraging Argo data, ? employed stationary
84 Gaussian process regression to compute temperature fields, estimating covariance parameters
85 through local moving-window maximum-likelihood estimation. They have enhanced uncertainty

86 estimates by recognizing the non-Gaussian and non-stationary nature of the temperature field. ?
87 introduced PROCAST, an efficient system that probabilistically predicts global mean air and sea
88 surface temperature variations, utilizing insights from CMIP5 simulations. Notably, data-driven
89 methods offer the advantage of providing physically meaningful error estimates and satisfactory
90 probabilistic forecasts. Additionally, they entail relatively lower numerical costs compared with
91 OGCMs.

92 In this current study, the Analog forecasting method (?) was combined with ensemble Data
93 Assimilation techniques, forming the AnDA method, as previously introduced by ? and Lguensat
94 et al. (2017). This method has already shown promising results in previous oceanic applications,
95 such as the successful reconstruction of the sea surface height in the Gulf of Mexico (Zhen
96 et al. 2020). What distinguishes this approach from traditional data assimilation schemes is its
97 integration of analog forecasting. By statistically learning the dynamics of the system from a
98 comprehensive catalog of analogs - encompassing a wide range of potential system states along
99 with their subsequent states - the method is adept at forecasting any given state. Consequently, it
100 can confidently estimate the evolution of a state at a given time toward the subsequent time step
101 when the conditions allow predictability, or, alternatively, remain uncertain and stagnant when the
102 state is less predictable.

103 The aim of the study was to acquire knowledge by constructing a catalog of analogs from a
104 gridded analysis over the well-sampled Argo period and to train AnDA to extrapolate temperature
105 observations over a less well-sampled period (prior to Argo). This approach enabled the assessment
106 of new information on temperature variability, both at the surface and at depth, during a period when
107 *in situ* measurements were lacking. To meet the challenge of processing observations that are sparse
108 in time and three-dimensional in space, such as a large set of hydrographic profiles, the method was
109 combined with the Reduced-Space Interpolation introduced by Kaplan et al. (1997). Specifically,
110 the parameters interpolated by AnDA are the temporal coefficients of the spatial-temporal signals
111 projected into a reduced set of three-dimensional Empirical Orthogonal Functions (EOFs), which
112 are built using the learning period. Utilizing EOFs offers the advantage of extracting dominant
113 patterns of variability, smoothing the analyzed field, and reducing the dimensionality of the system.
114 ? have demonstrated the effectiveness of using EOFs for reconstruction purposes in the Pacific.
115 The integration of space reduction and AnDA is termed RedAnDA.

116 The test region selected for implementing the ocean temperature reconstruction was the tropical
117 Pacific. This basin exhibits subsurface OHC variability, predominantly driven by the interannual
118 to decadal ENSO climate mode (e.g., ?). The ENSO phases redistribute heat throughout the basin,
119 with El Niño events resulting in warming of the upper 100 meters and cooling of the subsequent
120 400 meters, whereas La Niña events exhibit about the reverse pattern. This well-documented
121 interannual variability has been extensively studied (e.g., ???). During the Argo period, i.e., the
122 learning period for the catalog, 6 El Niño and 5 La Niña events were observed.

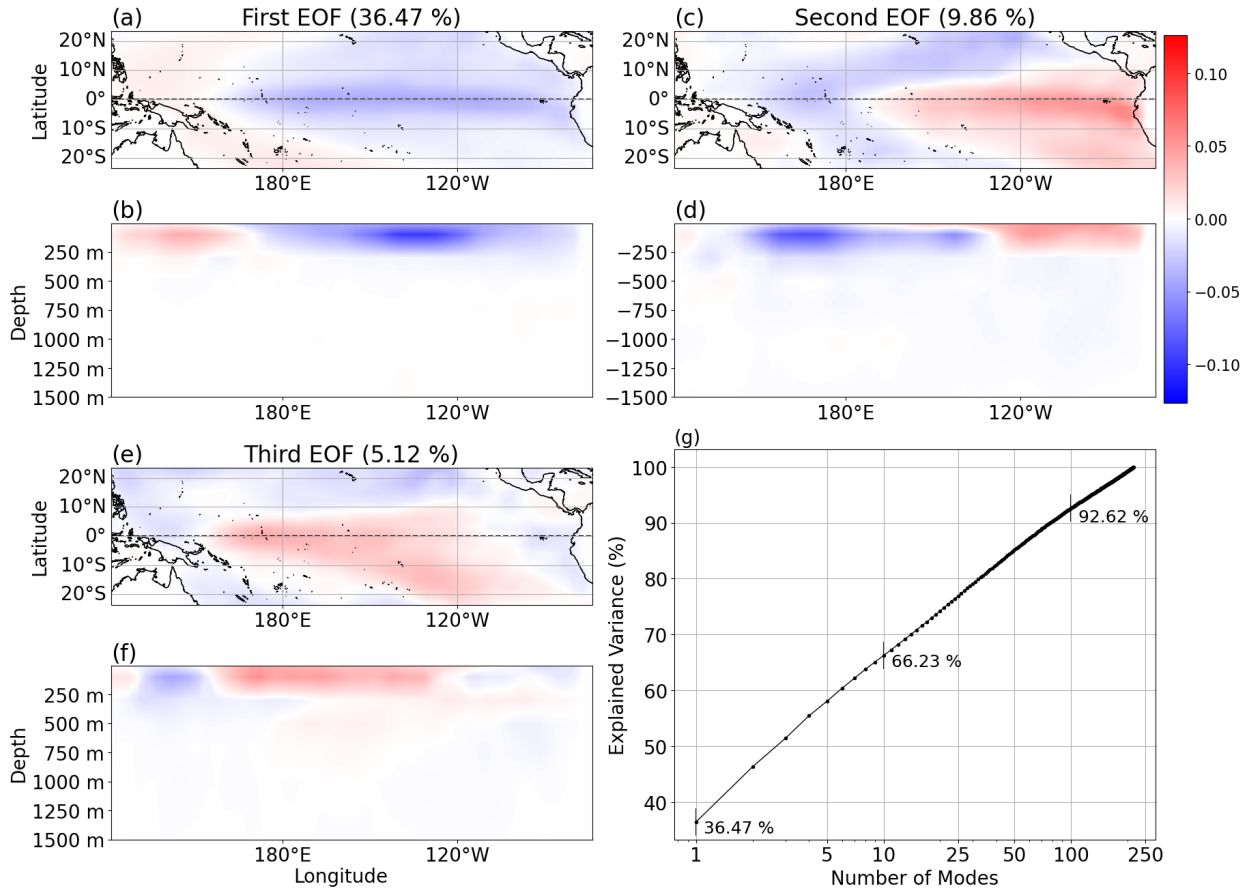
123 The paper is organized as follows. Firstly, the various procedures of the RedAnDA method
124 are detailed. The third section presents the different datasets used in this study. The fourth
125 section presents the validation of the method using an Observing System Simulation Experiment
126 (OSSE) approach. Finally, the application of RedAnDA to real observations and its comparison
127 are analyzed.

128 2. The RedAnDA Method

129 The RedAnDA method is a combination of three established methodologies: i) Space Reduction
130 utilizing Empirical Orthogonal Functions, ii) Analog Prediction, and iii) Data Assimilation.

131 After removing the monthly seasonal cycle, the three-dimensional temperature anomaly field
132 $\mathbf{T}(t)$ from the learning period, in the tropical Pacific Ocean, is considered. Empirical Orthogonal
133 Function (EOF) decomposition is then applied to reduce the dimension of the temperature field.
134 This method concentrates the variance into a reduced number of functions and relegates insignif-
135 icant portions to noise (Emery and Thomson 2001). This space reduction facilitates the Analog
136 Prediction, as discussed later in this section. By retaining the L dominant EOFs, the noise is
137 efficiently filtered out at the expense of a relatively small truncation error ϵ^r . The selection of L
138 involves a trade-off between system reduction and variance representation. Before the decompo-
139 sition, the prior weighting $\mathbf{w} \times \mathbf{T}(t)$, where \mathbf{w} represents the square root of the cosine of latitude,
140 ensures a uniform impact in terms of variance across different regions, accounting for their surface
141 area (Baldwin et al. 2009; Shea 2013).

146 The respective temporal coefficients of the EOFs, denoted $\alpha(t)$ set, are obtained by projecting
147 the spatial-temporal signals onto the set of eigenvectors. Variations in $\alpha(t)$ indicate changes in
148 the covariance patterns over time. Thus, the field approximation can be expressed as follows:



142 FIG. 1. First (a,c), second (b,d), and third (e,f) EOF computed from ISAS20 temperature at 5-meter depth and
 143 along an equatorial section. The explained variance for each of the three EOFs are given in the title of panels
 144 (a), (b), and (e). (g) Cumulative explained variance for 252 modes, as the number of EOFs was increased, with
 145 the x-axis presented on a log-scale. The values of the explained variance for 1, 10, and 100 modes are indicated.

149

$$\mathbf{w} \times \mathbf{T} = \alpha(t) \mathbf{EOF} + \epsilon^r(t) \quad (1)$$

150 The EOF decomposition was applied to the tropical Pacific using the learning period. As noted
 151 by Kaplan et al. (1997), the variance spectrum of the EOFs could decrease too steeply due to the
 152 initial coarsening of the field, leading to a strong constraint and low error estimates. To address
 153 this issue, energy redistribution was performed according to equation (19) in Kaplan et al. (1997).
 154 The patterns of variability are associated with the tongue in the equatorial band, extending from the
 155 east to the central Pacific and the warm pool in the western Pacific (Fig. 1 a and 1c). The transect

(Fig. 1c) displayed a concentration of correlation in the subsurface layers, primarily within the first 200 m. Such patterns are typical of Eastern Pacific El Niño events (Kao and Yu 2009). The second mode captures the information on the variability of Central Pacific El Niños (Kao and Yu 2009) (Fig. 1 b and d). However, owing to the orthogonality condition of the EOFs, the physical interpretation attributed to each mode was partially flawed, potentially leading to an overlap of phenomena. Thus, the combination of these two leading EOFs is required to express the diversity of El Niño events (Trenberth and Stepaniak 2001). The following modes held lesser physical significance. The third mode was not easily interpreted (Fig. 1 e and f). The following modes explained on average 0.23% of the initial variance, (Fig. 1 g). Nonetheless, they remained valuable for consistently conveying information from observations to unsampled locations, even at depth. A fundamental assumption of our study was that these correlation distributions, represented by the eigenvectors, were robust and remained applicable in the historical context.

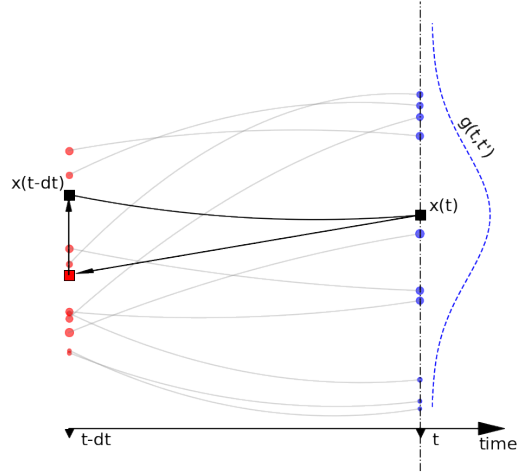
Second, RedAnDA differs from classical Data Assimilation schemes in that it replaces an explicit dynamical model with the Analog Backward Prediction, also known as Analog Backcast. It temporally propagates the interpolated information, offering informative *a priori* estimates of the ocean state, which are subsequently updated using available observations at corresponding times.

The use of analog prediction as a dynamical model offers several advantages, including eliminating the need to run Ocean General Circulation Models and reducing computational costs. However, it requires a dataset representative of the system’s dynamics, ideally extensive enough to encompass all potential system states. This dataset, termed a catalog, includes pairs of state vectors separated by the same time offset dt , with the first element known as the analog and the second as its temporal predecessor. The catalog can be constructed from observational data or numerical simulation results. In this study, the analogs consisted of temporal coefficients $\alpha(t)$ and their predecessors $\alpha(t - dt)$ over the learning period, i.e., the Argo period.

A classical approach to seek analogs of a current state is by searching for its k nearest neighbors in the state space defined by a Gaussian kernel:

$$g(t, t') = \exp(-\delta^2 |\alpha(t) - \alpha(t')|^2), \quad (2)$$

where t represents the time when the backcast is performed and t' is another time for which a record exists in the catalog. $|\cdot|$ denotes the Euclidean distance, and δ is set as the inverse of



180 FIG. 2. The process of analog backcast prediction. The black square $x(t)$ represents the current state for which
 181 backward evolution is computed. The blue dots represent the analogs, and the red dots represent their temporal
 182 predecessors. Here, the size of the analog sample is $k = 10$. The size of each dot reflects the weighting according
 183 to similarity, as determined by the function $g(t, t')$ from equation 2, represented by the blue hatched line. The
 184 red square represents the first prediction, which results from weighted linear regression. Finally, the black square
 185 of $x(t - dt)$ represents the final result, drawn randomly around the first prediction.

190 the median of the Euclidean distances of the sample of the k nearest neighbors. The value of k
 191 was set to 170. This value could not be significantly increased due to limitations imposed by the
 192 catalog size n , as the quotient $k/n = 170/252$. However, this choice was expected to ease strong
 193 convergence of the linear regression in the Analog Backcast, albeit at the expense of increased
 194 computation time.

195 As the size of the state space, i.e., the number of EOFs, increases, the search for analogs
 196 rapidly becomes more difficult, a phenomenon known as the "Curse of Dimensionality" (?). This
 197 complexity makes the reduced space approach convenient in this case.

198 At time t , when the analog backcast \mathcal{A} is applied to obtain an estimate of $\alpha(t - dt)$, the k
 199 analogs of the current state $\alpha(t)$ and their predecessors are selected. Each pair of analog and
 200 predecessor is then weighted according to equation (2). Assuming that both selections belong
 201 to normal distributions, a linear fit is performed, following the weighted least squares method of

202 Cleveland (1979). The obtained linear regression is used to extend $\alpha(t)$ at $t - dt$, providing a first
 203 estimate of the backcast (see Fig. 2). The final backcast is drawn from the normal distribution, with
 204 the mean being the first estimate and the spread being the weighted covariance of the differences
 205 between the linearly extended analogs and their respective predecessors. This randomness reflects
 206 the uncertainties in the prediction of the analog method, which are partially controlled by the
 207 chaotic dynamics of the system. When it is observed that the linear fit efficiently links the most
 208 important analogs and their associated predecessors, the prediction is kept around its first estimate.
 209 Otherwise, the final random draw would lead to a more dispersed prediction around the first
 210 estimate. These variations in the dynamics of the system are reflected by changes in the persistence
 211 of α samples. This adaptive approach is an advantage over sequential interpolation techniques
 212 that determine temporal persistence mostly on *a priori* statistics. The uncertainty of the analog
 213 prediction is given by a Monte Carlo procedure, to determine the spread of the different trajectories
 214 accessible from a set of given states, relatively close in the state space.

215 EOFs and temperature climatology were calculated over the Argo period, represented by the
 216 ISAS20 (In Situ Analysis System) temperature field (Gaillard et al. 2016). Associated $\alpha(t)$ values
 217 between 2002 and 2020 were used in the catalog of RedAnDA.

218 The reduced set of $\alpha(t)$ between 1950 and 2000 was then reconstructed for each month, by
 219 integrating the available *in situ* observations at time t with the Analog Backward Prediction \mathcal{A} ,
 220 from time $t + dt$ of the analysis, where dt represents one month. The assimilated profiles came
 221 from EN4.2.2 dataset (Good et al. 2013).

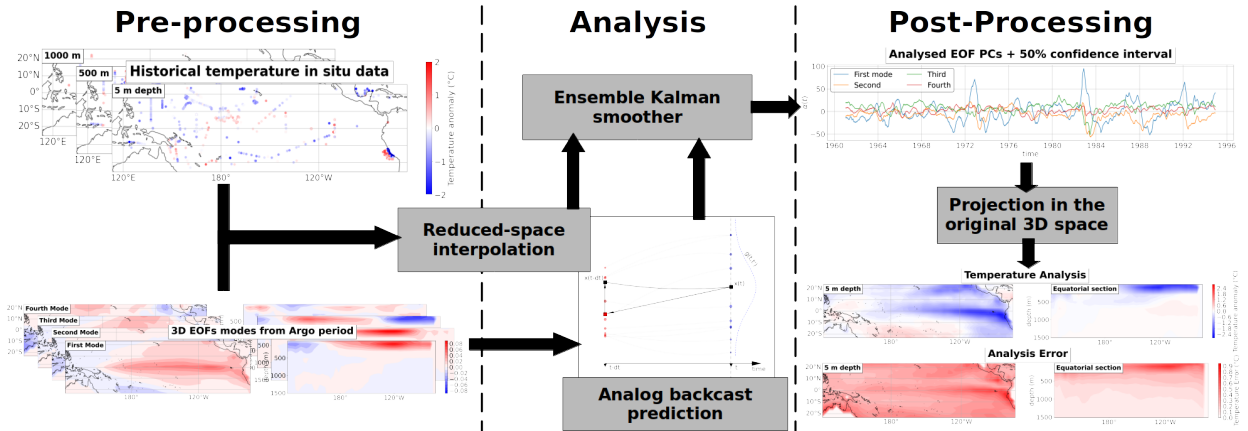
222 In this study, sequential RedAnDA was executed in a backward temporal fashion. By starting
 223 from the state $\alpha(t + dt)$, $\alpha(t)$ is determined through the application of the Analog Backcast \mathcal{A} .
 224 A notable distinction from forward-time analysis was observed during the initialization phase,
 225 wherein the initial α values were assessed based on varying quantities of available observations.
 226 This approach allowed for the transmission of comprehensive information from recent years
 227 backward in time, contrasting with the conventional forward methodology which propagates
 228 sparser historical data into the present. The RedAnDA scheme is as follows:

229

$$\alpha(t) = \mathcal{A}\{\alpha(t + dt), \epsilon^m(t + dt)\}, \quad (3)$$

$$\mathbf{y}(t) = \mathcal{H}(t) \boldsymbol{\alpha}(t) + \boldsymbol{\epsilon}^o(t). \quad (4)$$

231 The Analog Backcast \mathcal{A} uncertainty is denoted as $\boldsymbol{\epsilon}^m$. The observed data are represented by \mathbf{y} ,
 232 while $\mathcal{H}(t)$ refers to the linear interpolation operator linking the EOF grid points to the closest
 233 observation points available at time t . The uncertainty associated with the observations is denoted
 234 as $\boldsymbol{\epsilon}^o(t)$.



235 FIG. 3. RedAnDA analysis in the middle column, initiated by an initial temperature product drawn from a
 236 learning period and pre-processing of observation data shown in the left column. The post-analysis steps lead to
 237 the final gridded temperature product, in the right column.

238 The Reduced Space Interpolation method proposed by Kaplan et al. (1997) is consequently
 239 employed to define $\mathcal{H}(t)$ and $\boldsymbol{\epsilon}^o(t)$, with detailed elaboration provided later in the section.

240 The analysis proceeds in accordance with the Ensemble Kalman Smoother, augmented with the
 241 Analog Backcast as depicted in Fig. 1 of ?. Within the Ensemble variant of the Kalman filter,
 242 $\boldsymbol{\alpha}(t)$ and $\mathbf{y}(t)$ are treated as stochastic variables, following normal law distributions whose mean
 243 value and variance are determined via a Monte Carlo procedure. The variance serves to estimate
 244 the uncertainty associated with the variable.

245 Once $\boldsymbol{\alpha}(t)$ has been reconstructed for all desired time steps, the spatio-temporal temperature
 246 anomaly field is subsequently derived through re-projection into real space, utilizing the EOF-
 247 patterns. The whole RedAnDA procedure is summarized in figure 3. The Ensemble Kalman
 248 filter process is schematized in figure 2. The ensemble size in the Kalman Smoother was set to

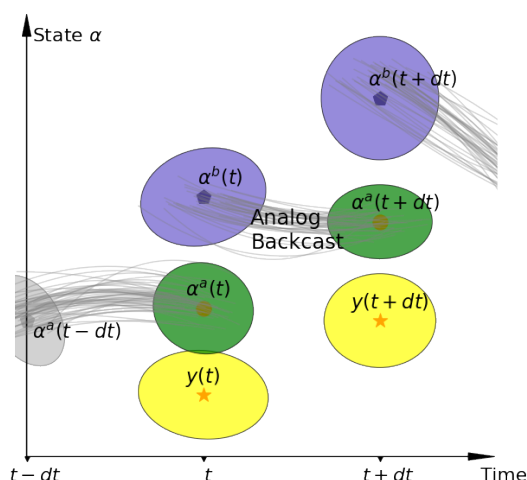
249 500 members to ensure a more reliable calculation of the mean and covariance at each time step of
250 the analysis. Increasing the ensemble size primarily impacted the analysis time.

251 Before applying RedAnDA to real observations, its efficiency is assessed through an OSSE (?).
252 We apply it to the monthly output from the Oceanic Chaos – Impacts, Structure, Predictability
253 (OCCIPUT) project (Penduff et al. 2014; Bessières et al. 2017). The simulation spans the period
254 from 1961 to 2015. The 408 months between 1961 and 1994 are selected for reconstruction, with
255 synthetic observations being the only source of information during this period. The 252 months over
256 1995-2015, for which full grid temperature maps are available, provide analogs and predecessors
257 for the Analog prediction. The full grid maps of the reconstruction period are further used to
258 qualify the target truth. The way these periods are split is designed to mimic the real situation
259 where we have a recent and short Argo period with good sampling and a longer pre-Argo period,
260 for which fewer observations are available.

261 In the OSSE, for comparison with the RedAnDA method, a standard Optimal Interpolation
262 method is also implemented, following Good et al. (2013). We call it OI. It reconstructs monthly
263 temperature maps sequentially, from 1961 to 1994, using an *a priori* monthly climatology built
264 during the learning period, to which $\sim 82\%$ of the previous month's analyzed anomaly is added.

265 To lighten the calculations and focus on the large spatial scale, all gridded products, for the OSSE
266 and for real reconstruction, are coarsened down to $3^\circ \times 3^\circ$ by horizontal averaging. Depth levels are
267 restricted to 5, 100, 300, 500, 700, 1000, and 1500 m. The tropical Pacific is delimited here between
268 120°E and 70°W , 25°S and 25°N . EOFs are calculated from the resulting temperature field. For
269 the space reduction, the number of retained EOFs was fixed at $L=10$, ensuring that approximately
270 65% of the total variance of the initial true temperature field was explained. However, the primary
271 EOFs primarily accounted for the high variability of the surface. To capture variability at depth,
272 additional modes were required. With $L=10$, temperature signals at depths of 5, 100, 300, 500,
273 700, 1000, and 1500 meters were reconstructed by approximately 67% , 73% , 44% , 27% , 22% ,
274 23% , and 20% , respectively.

282 In practice, $\epsilon^o(t)$ from equation (4) was computed by combining ϵ^r , representing the uncertainty
283 of the truncated EOFs, and an empirically derived representativeness standard deviation. This
284 standard deviation reflected the square root of the variance of all EN4.2.2 profiles found in large
285 horizontal boxes, at each depth. For the real observation uncertainty, $\epsilon^o(t)$ also encompasses



275 FIG. 4. Schematic diagram of the Kalman Filter process within the Data Assimilation method. Beginning with
 276 the analyzed state α^a resulting from the assimilation of observations (depicted in green), the analog backcast
 277 was executed to provide an estimate of the state at the previous time step α^b (represented in blue). Subsequently,
 278 this estimated state is updated using the *in situ* observations y (depicted in yellow), and this iterative process
 279 continues until the entire period under consideration has been processed. Each variable is depicted with its
 280 respective uncertainty, symbolized by the area of the colored shapes, which is correlated with the spread of the
 281 ensemble member trajectories (illustrated by the gray lines).

286 instrumental error, as assessed following ?. This error ranged from 0.3°C for MBTs and moorings
 287 to 0.002°C for the most precise CTDs.

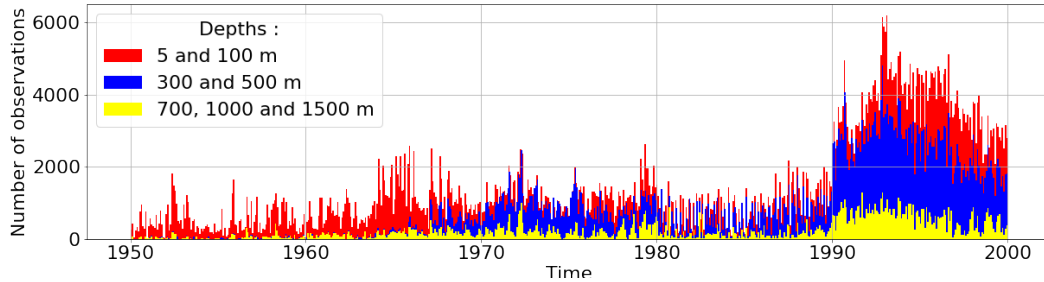
288 While the sampling of observations was generally adequate to support monthly analysis, during
 289 the early period of the time series, the availability of profiles could be limited, leading to potentially
 290 significant spurious fluctuations at large scales. To address this issue, a procedure was implemented
 291 to automatically extend the temporal window for assimilating observations. The condition for
 292 this extension was to have a minimum of 5 000 observations before resuming the analysis. This
 293 threshold was chosen because its magnitude was similar to the average number of monthly available
 294 observations in the most recent years of the analysis. At each time step, care was taken to ensure
 295 that the temporal window did not exceed the most energetic time periods of the selected EOFs,
 296 thereby preventing damping of the interpolated result.

297 3. Data

298 The ISAS20 (In Situ Analysis System) gridded product was chosen to represent the Argo period
299 (Gaillard et al. 2016). This product provides monthly $0.5^{\circ} \times 0.5^{\circ}$ gridded temperature and salinity
300 fields, constructed using an Optimal Interpolation method. It encompasses 187 standard depth
301 levels distributed between 0 and 5500 meters and covers the period from 2002 to 2023. It
302 constitutes an analysis of *in situ* measurements, aiming to preserve the time and space sampling
303 capabilities of the Argo network (Kolodziejczyk et al. 2021). Therefore, two correlations scale
304 were used: 300 km and 4 times the deformation radius (Gaillard et al. 2016).

305 The assimilated profiles for reconstructing past periods came from the EN4.2.2 dataset (Good
306 et al. 2013). This dataset represents the latest version of the Met Office Hadley Centre's 'EN'
307 dataset series, offering quality-controlled potential temperature profiles from 1900 to the present
308 day. Prior to the Argo period, most profiles were sourced from the World Ocean Database 2018
309 (WOD18) (Boyer et al. 2006), and from the Global Temperature and Salinity Profile Program
310 (GTSP) (Sun et al. 2010). Notably, no Argo profiles were included in the latter dataset. In the
311 synthetic profiles of the OSSE as well as in EN4.2.2, only depths where the temperature observation
312 had a quality flag of 1 are retained. Additionally, since these depths varied from profile to profile,
313 a vertical linear interpolation was conducted to obtain the profiles at the target reconstruction grid
314 levels. No extrapolation was performed. To filter out high frequencies in observations from fixed
315 tropical moorings (TAO/TRITON array in the Pacific), it was decided to average the time series
316 over 15-day periods for each identified buoy. Super-profiles were then considered at the center of
317 the time slice, reducing the number of initial profiles from 832 703 to 648 786. In figure 5, the
318 number of available observations, between 1950 and 2000, appears to increase over time, with
319 a substantial improvement in sampling in the nineties. Two minima are found in the fifties and
320 eighties, suggesting that reconstruction during these periods may have been less constrained.

325 In the OSSE framework, the Oceanic Chaos – Impacts, Structure, Predictability (OCCIPUT)
326 project was utilized (Penduff et al. 2014; Bessières et al. 2017). The OCCIPUT project offers an
327 ensemble of 50 realizations of a global $1/4^{\circ}$ ocean/sea ice simulation, based on the NEMO 3.5
328 model (Penduff et al. 2014; Bessières et al. 2017), alongside synthetic profiles extracted at specific
329 times and locations corresponding to the historical database. For this study, only one member of



321 FIG. 5. Monthly histogram of available EN4.2.2 temperature observations in the tropical Pacific, depending
 322 on the depth, between 1950 and 2000. In red, observations at 5 and 100 meters depth are counted. In blue,
 323 observations at 300 and 500 meters depth are counted. The number of observations at 700, 1000 and 1500 meters
 324 depth is represented in yellow.

330 the model output and synthetic profiles were used. The simulation spans the period from 1961 to
 331 2015, employing the DFS atmospheric forcing over the same period (Dussin et al. 2016).

332 To compare RedAnDA’s temperature reconstruction with state-of-the-art reanalysis products
 333 since the 1950s, five products were introduced. Three of these are optimal interpolation products,
 334 while the other two are derived from data assimilation with numerical models:

- 335 • ISHII: Version 6.4.0 of a historical objective analysis of temperature data provided by the
 336 Japan Marine Science and Technology Center. ISHII utilizes *in situ* observations from the
 337 World Ocean Database 2005 (WOD05), the Global Temperature and Salinity Profile Program
 338 (GTSP), Centennial *in situ* Observation Based Estimates (COBE), and ARGO buoys (?). It
 339 compiles monthly $1^\circ \times 1^\circ$ temperature fields on 24 depth levels.
- 340 • EN4: A gridded product from the Hadley Centre of the UK Meteorological Office. EN4
 341 employs an optimal interpolation technique, with *in situ* profiles from the EN4.2.2 dataset
 342 analyzed using a smoothed 1971–2000 field from the EN2 version (Good et al. 2013). It
 343 provides monthly temperature fields from 1900 to the present day over 42 depth levels.
- 344 • IAP product: Provided by Ensemble Optimal Interpolation of the CODC-GOSD dataset, it
 345 utilizes CMIP5 models to reconstruct global and gridded temperature at a monthly period
 346 from 1940 to the present day, at a spatial resolution of $1^\circ \times 1^\circ$ in the upper 2000 m, over 41
 347 levels (??) .

- 348 • ESTOC dataset: Developed by the Japan Agency for Marine-Earth Science and Technol-
349 ogy (JAMSTEC) and Kyoto University, ESTOC utilizes a four-dimensional variational data
350 assimilation method with the GFDL Modular Ocean Model (MOM3). It assimilates quality-
351 controlled EN3 *in situ* profiles and sea-surface dynamic-height data from Aviso altimetry
352 products (??). The $1^\circ \times 1^\circ$ monthly result covers the period from 1956 to 2009 over 46 vertical
353 levels.
- 354 • ORAS4: Provided by the Copernicus Climate Change Service, ORAS4 combines model
355 data with observations using the NEMO model and the 3D-Var assimilation system called
356 NEMOVAR (?). It assimilates *in situ* profiles from the EN3 dataset, sea-surface temperature
357 from the OIv2 product, and altimeter sea level anomalies from the AVISO dataset. ORAS4
358 covers the period from 1959 to the present day on 42 depth levels and with a horizontal spatial
359 resolution of $1^\circ \times 1^\circ$.

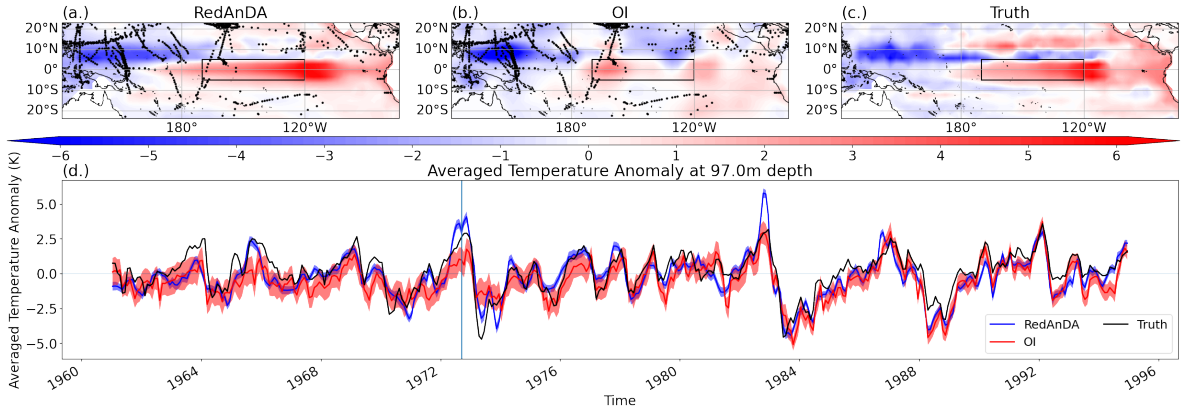
360 4. Validation using numerical simulations and OSSE

361 The tropical Pacific temperature fields were reconstructed within the OSSE framework, with
362 the assimilation of synthetic observations available between 1961 and 1994. The 1995-2015
363 OCCIPUT model (Penduff et al. 2014; Bessières et al. 2017) fields served as a learning period for
364 the Analog backcast.

365 Our validation initially focused on selected depths (around 100 m, 500 m, and 1000 m), analyzing
366 the temporal evolution of temperature anomalies averaged over specific sub-regions chosen to
367 highlight the key features of both RedAnDA and OI products, as well as the model truth.

368 We first present the reconstructions in the tropical Pacific at 100-m depth level (Fig. 6a, b, and c).
369 During the 1973 El Niño, as identified in the Niño 3.4 time series (Fig. 6d), the temperature field
370 appeared consistently reconstructed with RedAnDA (Fig. 6a and c). However, the OI reconstruction
371 lacked the necessary observations to successfully capture the spatial distribution of the warm and
372 cold anomalous patterns of the El Niño event (Fig. 6b). Despite persistence from the initial guess,
373 which allowed the emergence of a warm signal around 160°E , the region remained overall too
374 cold in the absence of surface observations. In contrast, the cold anomaly in the North Equatorial
375 Countercurrent around $5\text{-}10^\circ\text{N}$ was satisfactorily reconstructed with RedAnDA (Fig. 6a and c).

376 Due to an insufficient number of observations and inaccurate *a priori* correlation scales, OI did not
 377 reproduce such fine structure (Fig. 6b and c).

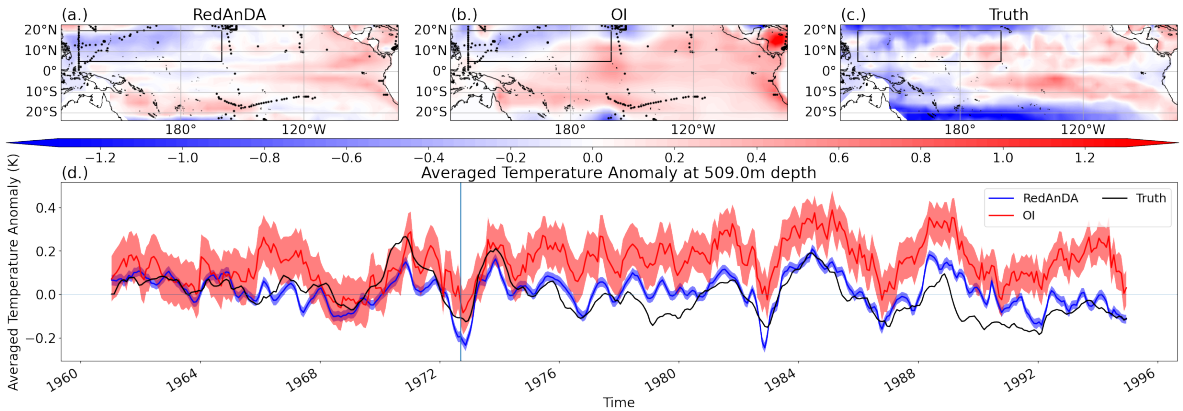


378 FIG. 6. (a, b, c) Temperature anomalies at around 100-m depth for RedAnDA, OI, and OSSE truth in September
 379 1972. The black dots indicate synthetic observations available at the given depth and time. The black box in
 380 panels a, b, and c shows the region where the temperature is calculated to obtain a pseudo-Niño 3.4 index.
 381 (d) Temporal variations in this average for RedAnDA, OI, and OSSE truth. RedAnDA and EN4 provide 50%
 382 confidence intervals. The blue vertical line indicates the selected month for the top snapshots.

383 The time series of the averaged temperature anomaly in the Niño 3.4 box, a region where an SST
 384 anomaly index is typically calculated, indicated that both methods performed better in the recent
 385 past, from 1974 to 1994, when more observations were available (Fig.5 and 6d). In this selected
 386 region, the temperature anomaly was strongly positive during El Niño events, and most of them
 387 were successfully identified in the years: 1963, 1965, 1968, 1969, 1972, 1976, 1977, 1979, 1983,
 388 1987, 1990, 1992, 1993, and 1994. Interestingly, in the early period, the El Niño events of 1963,
 389 1965, 1972, and 1976 were systematically underestimated by about 1°C by the OI reconstruction.
 390 Similarly, most La Niña events were identified in both products (1962, 1964, 1971, 1973, 1974,
 391 1975, 1984, and 1988), but tended to be overestimated by 1-2°C, as in 1971, 1974, and 1988.
 392 Occasionally, such biases could have led to an overestimation of temperature, as in 1972 and 1983.

393 At 500-m depth, the true signal exhibited low-frequency variability that was scarcely detected
 394 by both OI and RedAnDA (Fig. 7a, b and c). This was particularly evident outside the equatorial
 395 zone, where a region with strong cold anomalies of approximately -1.2°C extended along latitudes
 396 20°N and 20°S . From the Coral Sea around 18°S and 158°E to the eastern boundary of the Pacific,
 397 both reconstructions failed to capture the extent of the cold anomaly (Fig. 7a, b, and c). Although

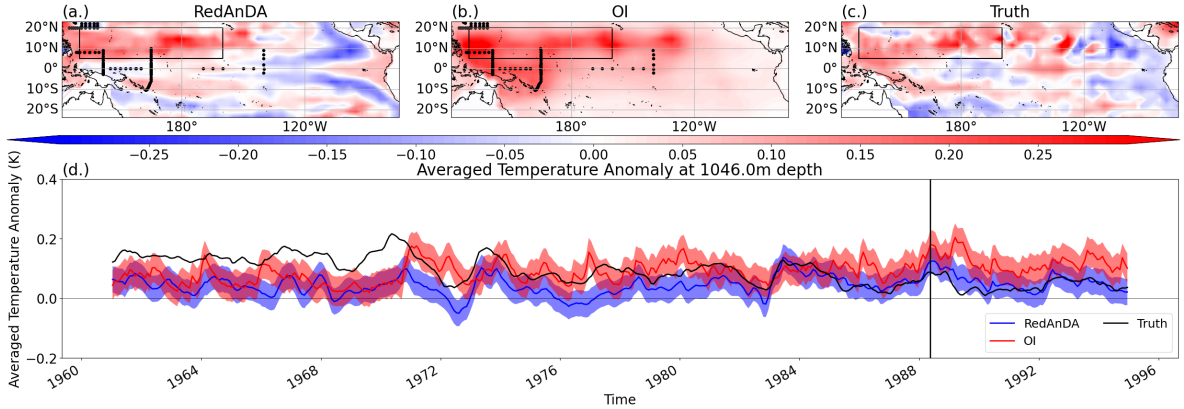
398 RedAnDA and OI reconstructed the eastern warm and northwestern cold regions, they did so with
 399 an intensity underestimated by approximately 0.2°C . Additionally, in the eastern tropics, the La
 400 Niña pattern at 500-m depth in RedAnDA was more realistic compared to that in OI. For instance,
 401 RedAnDA was able to identify the equatorial colder temperatures near the South American coast.



402 FIG. 7. (a, b, c) Temperature anomalies around 500 m depth for RedAnDA, OI, and OSSE truth in September
 403 1972. Black dots indicate synthetic observations available at the given depth and time. The black box in panels
 404 a, b, and c, now located between 5° - 20°N , 130°E - 160°W , shows the region where the temperature is calculated to
 405 estimate the signature of the westward North Equatorial Subsurface Current (NESC) (?). (d) Temporal variations
 406 in this average for RedAnDA, OI, and OSSE truth. RedAnDA and OI provide 50% confidence intervals. The
 407 blue vertical line indicates the selected month for the top snapshots.

408 The temporal variations illustrated RedAnDA's capability to reconstruct the averaged variability
 409 in the selected box in the northwest of the domain, between 5° - 20°N and 130°E - 160°W (Fig. 7c
 410 and d). The choice of the box was made to capture the highest regional variability at 500-m depth,
 411 which differed from the area at 5 m depth. This variability is associated with the North Equatorial
 412 Subsurface Current (NESC), which varies interannually along with ENSO. It flows stronger during
 413 La Niña events and weaker during El Niño (?). However, some variations appeared excessively
 414 pronounced in the RedAnDA product, such as in 1967 with a difference of 0.1°C , or in recent
 415 years, while others were underrepresented with similar magnitude, such as in 1971, around 1980,
 416 and between 1989 and 1992 (Fig. 7d). In contrast, OI adequately represented the period 1968
 417 to 1974 and exhibited variations that correlated with RedAnDA most of the time. However, it
 418 suffered from a drift that led to an overestimation of temperature by approximately 0.2°C in the

419 selected region over time. Moreover, the OI signal was notably noisy, due to the irregular sampling
 420 of observations.



421 FIG. 8. (a, b, c) Temperature anomalies around 1000 m depth for RedAnDA, OI, and OSSE truth in May 1988.
 422 Black dots indicate synthetic observations available at the given depth and time. The black box in panels a, b, and
 423 c, between 5-20°N, 130°E-160°W, shows the region where the temperature is averaged to assess the variability
 424 associated with the North equatorial deep jets (???). (d) Temporal variations in this average for RedAnDA,
 425 OI, and OSSE truth. RedAnDA and OI provide 50% confidence intervals. The blue vertical line indicates the
 426 selected month for the top snapshots.

427 Around 1000 m depth, the number of observations diminished significantly (Fig. 5 and 8). It
 428 impacted the OI reconstruction, particularly east of 120°W (Fig. 8b). In the northwest tropical
 429 Pacific, the OI captured the warm zonal pattern along 15°N with an anomaly of 0.25°C (Fig. 8b
 430 and c). In contrast, RedAnDA extrapolated more signal from analogs, aided by three-dimensional
 431 EOFs connecting the surface to the depths (Fig. 8a). Although its reconstruction may have been
 432 insufficient to represent small-scale structures. From 120°E to 120°W in the northern tropics and
 433 between 150°E to 100°W in the south, RedAnDA successfully retrieved the warm extensions, along
 434 with cold regions of approximately -0.2°C east of 120°W along the coast. Along the equator, the
 435 most significant differences between RedAnDA and Truth were observed, with dissimilar warm
 436 and cold patterns (Fig. 8a and c).

437 Considering the time series in the selected box, OI had difficulties capturing the most significant
 438 temperature variance (Fig. 8d), exhibiting high-frequency oscillations not coherent with the time
 439 series derived from the Truth. This is likely attributed to sparse and uneven distribution in data
 440 sampling. Initially, in 1961, OI closely aligned with RedAnDA temperature estimates, but then

441 diverged from model Truth and RedAnDA by up to 0.1°C, as persistence introduced a drift. Its
442 correction was impeded by either an insufficient number of observations or an overemphasis on
443 background signals.

444 RedAnDA's temperature time series correlated more closely with the truth, especially after 1983.
445 Before 1983, the oscillations were recovered, but often too weakly and biased, as the true signal
446 presented a drift. Before 1970, the difference due to the drift was greatest, and RedAnDA's
447 reconstructed oscillations correlated less closely with the truth, with a difference of 0.05°C in
448 1961.

449 To evaluate the quality of the uncertainty estimation, we computed the probability coverage. It is
450 evaluated as the proportion of the true signal within a given confidence interval of the reconstruction.
451 The confidence interval was determined by the normal quantile q , set to approximately 0.67 for a
452 50% confidence interval. This metric assessed the reliability of uncertainty estimates in capturing
453 differences between the true and reconstructed signals. A proportion close to 50% indicated
454 proper uncertainty estimation, while deviations from 50% indicated potential underestimation or
455 overestimation of uncertainty. For more details about this metric, see ?.

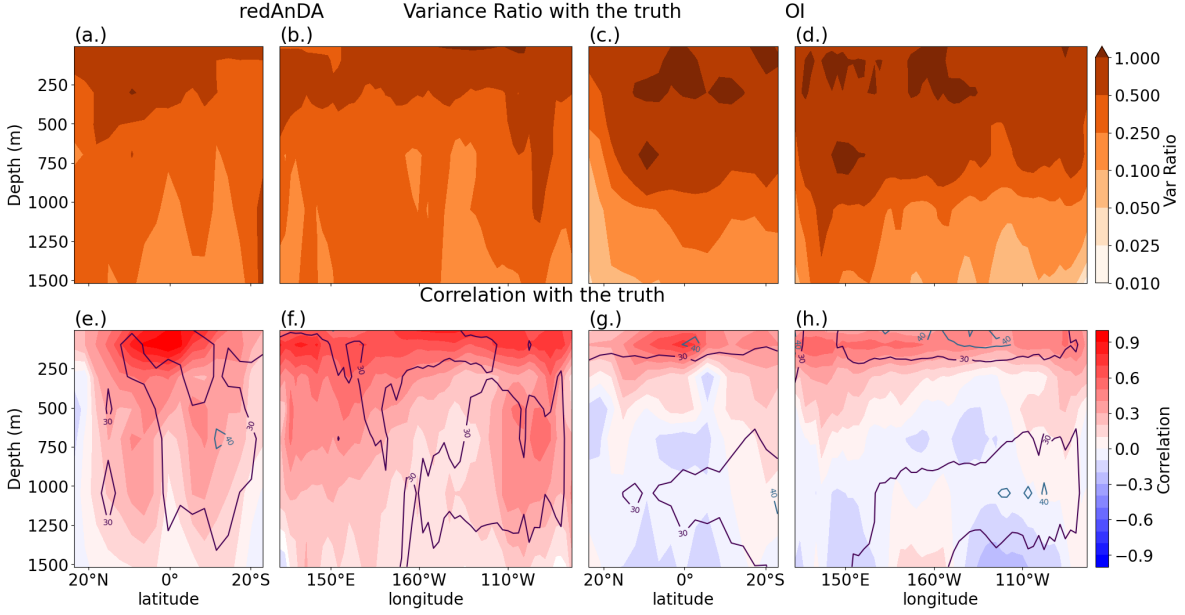
$$\text{Probability Coverage} = P\left[\mathbf{T}(t) - q \times \sigma \geq \mathbf{T}^{\text{true}}(t) \geq \mathbf{T}(t) + q \times \sigma\right], \quad (5)$$

456

457

458 Generally, both methods captured less than the desired amount of truth within their confidence
459 intervals (Fig. 9e, f, g and h). For OI, the uncertainty was basin-wide underestimated. Spatially,
460 only 30% of the truth was captured within the confidence interval over most of the domain where
461 *a priori* variances were the highest (Fig. 9c, d, g, and h). Below 300 m, estimates were the
462 least reliable, which could be attributed to observed drifts (Figs. 7d and 8d). At depth, in the
463 southern half of the basin, the reconstruction of anomalies was better, leading to higher inclusion
464 percentages, between 30-40% (Fig. 9g).

470 The RedAnDA uncertainty was linked to the space reduction, the sampling of observations, and
471 the dispersion of members during the backcast prediction. The confidence interval was small,
472 given the variability of the signal under consideration and the EN4 uncertainty (Figs. 6d and 7d).
473 Before 1970, the confidence interval was wider, but still failed to capture the true signal when



465 FIG. 9. Meridional and zonal mean sections depicting the ratio of reconstructed variance to true variance (first
 466 row) and the correlation of the two analysis products with the OSSE truth (second row). In the panels of the
 467 second row, purple and blue contours indicate the 30% and 40% values in probability coverage. For the variance
 468 ratio, note the logarithmic color scale. The two left columns represent the RedAnDA products, while the two
 469 right columns represent the OI products.

474 the drift was significant (Fig. 8d). Overall, the results were overconfident at the surface, as the
 475 inclusion of the truth within the confidence interval was less than 30 %. The quantification of
 476 uncertainty was better at depth (Fig. 9e and f).

477 Supplementary performance scores were calculated using several metrics. These metrics
 478 included the variance ratio and correlation defined as follows:

$$479 \text{ Variance Ratio} = \frac{\overline{\mathbf{T}(t)^2}}{\overline{\mathbf{T}^{\text{true}}(t)^2}}, \quad (6)$$

480
 481 The variance ratio compared the variance of the reconstructed signal \mathbf{T} , to the variance of the true
 482 signal \mathbf{T}^{true} . A variance ratio of 1 indicated that the reconstructed variability matched the true
 483 variability in energy.

$$\text{Correlation} = \frac{\overline{(\mathbf{T}(t) - \mathbf{T}^{\text{true}}(t))^2}}{\sqrt{\overline{\mathbf{T}(t)^2} \times \overline{\mathbf{T}^{\text{true}}(t)^2}}}, \quad (7)$$

The correlation measured the coherence between the true and reconstructed signals. It ranged from -1 and 1 , where -1 indicated opposite phase variations, 1 indicated similar variations, and 0 indicated no consistency between the signals.

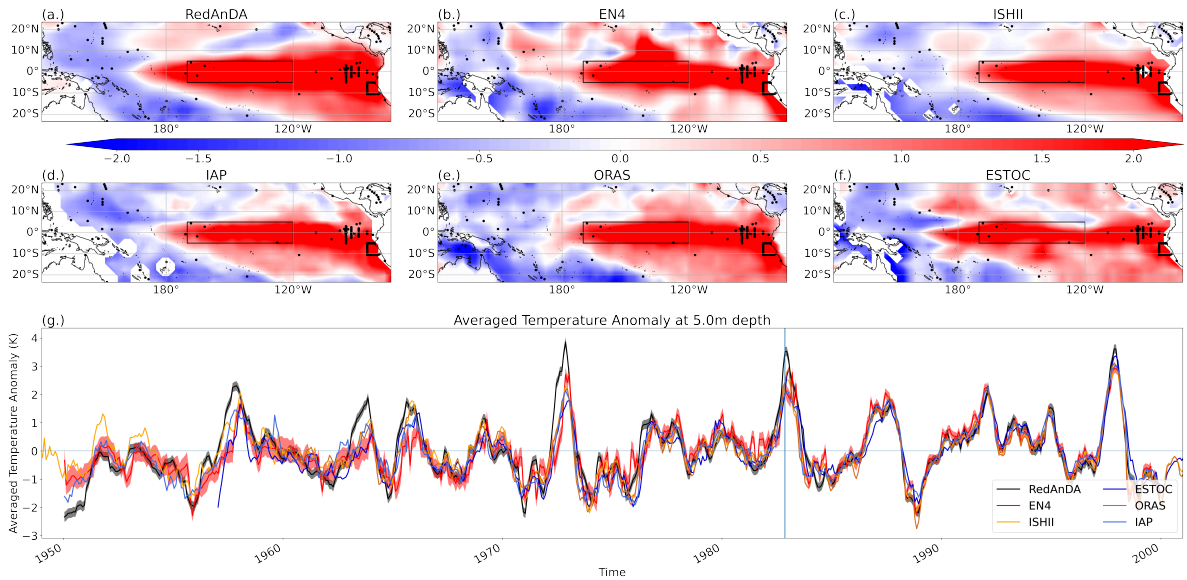
Both analyses demonstrated the capability to inject a satisfactory level of variability into the upper 300 m, where they also exhibited satisfactory correlations with the model Truth (Fig. 9a-h). In the warm pool to the west along the equator and in the warm extensions of El Niño, between 160°W and the eastern basin, the effectiveness of the RedAnDA reconstruction was notably evident (Fig. 9a, b, e, and f). This outcome was anticipated, given that the principal modes of EOFs were linked to the variability of these regions. In contrast, OI exhibited satisfactory variance ratios and correlations for reconstruction, primarily concentrated in regions with better sampling (Fig. 9c, d, g, and h).

At depths exceeding 300 meters, OI continued to inject information but displayed weak or negative correlations (ranging from 0 to -0.2) with the ground truth (Fig. 9c, d, g, and h). The signal reconstructed by RedAnDA exhibited more consistency with the truth, displaying positive correlation reaching up to 0.5 (Fig. 9a, b, e, and f). However, at the northern and southern boundaries, below 300 m, the reconstruction's accuracy was less evident, owing to drifts in the model truth that neither RedAnDA nor OI identified (Figs. 7d and 8d).

5. Results

The regional reconstruction of temperature was conducted using real observations spanning from 1950 to 2000 in the tropical Pacific. In this section, the performance of RedAnDA is compared with other temperature fields derived from various analyses.

Firstly, Sea Surface Temperature (SST) reconstruction maps were examined for the intense 1982 El Niño event (Fig. 10a-f), during which observation availability was not homogeneous. All reconstructions depicted the warm tongue extending 2°C to the east and the cold region to the west. Along 10°N , between 180°E and 150°W , the cold anomaly was detected by all methods.



504 FIG. 10. Temperature anomalies at 5-m depth in November 1982 for various datasets, including RedAnDA,
 505 EN4, ISHII, IAP, ORAS, and ESTOC. The upper two rows display snapshots of temperature anomalies for each
 506 dataset, with black dots indicating available observations at the given depth and time. The observation positions
 507 represented here may not precisely match those assimilated by analyses other than RedAnDA. The black box
 508 highlights the region where temperature was averaged to obtain a Niño 3.4 index. In the lower panel, the temporal
 509 variations of this average temperature anomaly are depicted. Both RedAnDA and EN4 provided 50% confidence
 510 intervals. The blue vertical line denotes the time selected for the upper snapshots. To enhance visibility, the
 511 RedAnDA curve is depicted in black.

520 However, around 120°W-15°N in the northern Tropical Pacific, there was disagreement among the
 521 reconstructions regarding the recovery of the cold anomalies (Fig. 10 a).

522 The Niño 3.4 index exhibited good agreement among all time series (Fig. 10 g). ENSO events
 523 in RedAnDA appeared to be slightly stronger (up to 1°C) compared to other products, although
 524 previous validation results in section 4 demonstrated its ability to accurately recover El Niño
 525 events.

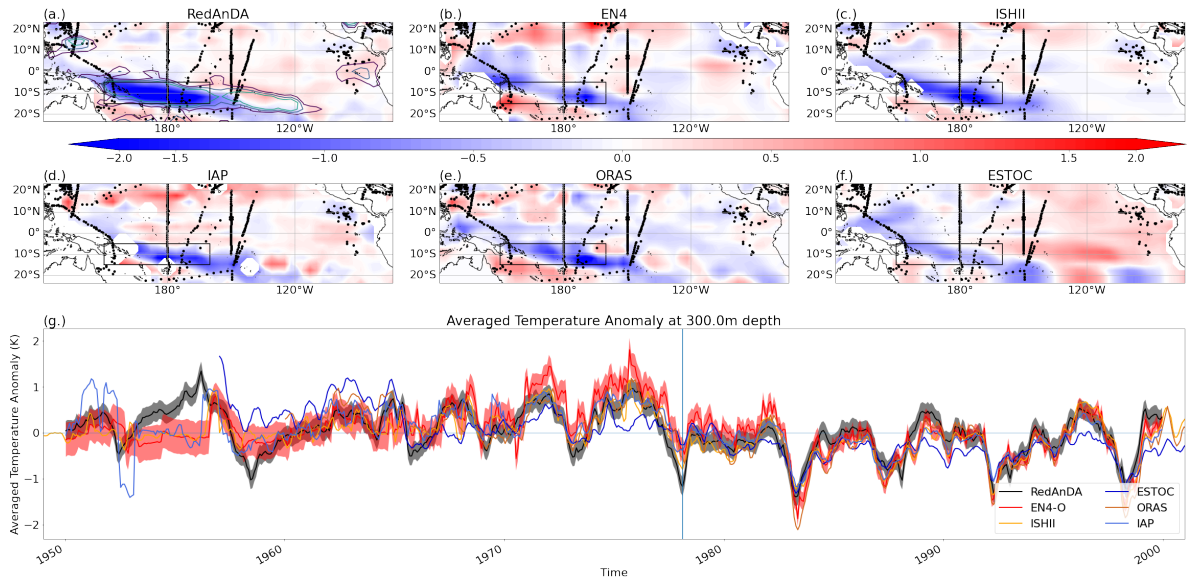
526 In the 1950s, intense cooling was observed in RedAnDA, which was not observed in the other
 527 analyses. Error estimates for both EN4 and RedAnDA decreased over time as the number of
 528 observations increased. The error interval for RedAnDA was notably thinner compared to EN4,
 529 potentially indicating an underestimated error in the reconstruction.

530 At a depth of 300 meters, in February 1978, the various methods yielded less consistent results
531 (Fig. 11a-f). In the southwest of the domain around 180°E, there was a consensus regarding
532 the presence of a cold anomaly pattern, approximately -2°C , with, on its southern and northern
533 flanks, warm anomalies ranging between 0.5 and 1°C . However, in the eastern Pacific, significant
534 disparities emerged despite the presence of some observations near the coast. Around 120°W,
535 between 15°S and 10°N, IAP, ESTOC, and RedAnDA appeared to agree on a warm anomaly
536 pattern whereas ORAS tended toward a colder anomaly (Fig. 11a, d, e, and f). In this region,
537 RedAnDA diverged from EN4 and ISHII, but these reconstructions, lacking sufficient observational
538 data, exhibited low energy (Fig. 11b and c).

539 In the box (Fig. 11a-f), located in the southwest of the study area, between 150°E - 160°W and
540 5 - 15 °S, temporal fluctuations relate in part to the Pacific South Equatorial Current (SEC). This
541 current undergoes interannual fluctuations in its intensity related to ENSO, as it flows stronger a
542 few months following an El Niño event and weaker after a La Niña event (?). Concerning the
543 estimation of this variability, after the 1970s, the different time series are aligned (Fig. 11 g). In
544 1979, RedAnDA reconstructed a significant cold event with a magnitude around 1°C , which was
545 less pronounced in other methods, estimating its magnitude at -0.5°C . This discrepancy arose from
546 the cold extension between 5°-15°S and 150°E to 120°W and the warm southern pattern centered
547 at 15°S and 165°E, which were larger and narrower, respectively, compared to other methods (Fig
548 11a-f).

549 A similar discrepancy occurred in 1959. In the 50s, RedAnDA assimilated more observations
550 over a wider time window (see section 2), resulting in a smoother analysis during this period than
551 in other analyses. Except for the 1959 event, the variations correlated with the highest peaks in
552 IAP. However, these were likely associated with spurious assimilation of sparse observations.

564 The average correlation of RedAnDA with the other products indicated overall agreement (Tab. 1).
565 ISHII and ORAS showed the closest similarity to RedAnDA, with both correlations at 0.61, while
566 ESTOC showed the least similarity with a value of 0.47. These close scores demonstrated that
567 RedAnDA aligned with the state-of-the-art reconstructions. However, the existence of dissimilarity
568 suggested that RedAnDA still provided original results, not yet obtained with any of the alternative
569 methods. Regarding the results from the OSSE, the original reconstruction of RedAnDA in the



553 FIG. 11. Temperature anomalies at 300m depth in February 1978, including RedAnDA, EN4, ISHII, IAP,
 554 ORAS, and ESTOC. The upper two rows display snapshots of temperature anomalies for each dataset, with black
 555 dots indicating available observations at the given depth and time. The observation positions represented here
 556 may not precisely match those assimilated by analyses other than RedAnDA. The contours in (a) delimit the 0.35,
 557 0.45 and 0.55°C temperature standard deviations to highlight the regions of high variability. The black box,
 558 now between 150°E - 160°W and 5 - 15 °S, has changed to highlight the southwestern region where temperature
 559 was averaged to visualize variability partly associated with the Pacific South Equatorial Current (SEC) (??).
 560 In the lower panel, the temporal variations of this average temperature anomaly are depicted. Both RedAnDA
 561 and EN4 provide their 50% confidence intervals. To enhance visibility, only these are shown. The blue vertical
 562 line denotes the time selected for the upper snapshots. To enhance visibility, the RedAnDA curve is depicted in
 563 black.

570 upper 500 m depth may be worth consideration as it may have reconstructed components of the
 571 variability that were undetected from other datasets.

575 6. Discussion & Conclusion

576 In conclusion, the RedAndA method represents a pioneering approach by integrating Analog
 577 backward prediction Data Assimilation with reduced-space analysis. It demonstrated successful
 578 reconstruction of the monthly three-dimensional temperature field from sparse and randomly
 579 distributed *in situ* profiles in the Pacific Ocean. In the OSSE, RedAnDA outperformed Optimal

	Correlation
ISHII	0.61
EN4	0.52
IAP	0.57
ORAS	0.61
ESTOC	0.47

572 TABLE 1. Comparative analysis of the RedAnDA reconstruction with reconstructions from ISHII, EN4, IAP,
573 ORAS, and ESTOC, listed in rows from first to fifth, respectively. The scores represent the Correlation. Each
574 comparison is conducted over the period common to both reconstructions.

580 Interpolation (OI) in terms of truth retrieval. Moreover, it exhibited agreement with alternative
581 reconstructions such as ISHII, IAP, ORAS4, and ESTOC, while inferring variability where other
582 methods lacked information and avoiding spurious results.

583 RedAnDA’s advancements relied on its ability to extrapolate to regions with poor sampling, par-
584 ticularly at depth, facilitated by 3D EOFs that correlated the different vertical levels. Additionally,
585 the analog backcast provides statistical temporal predictions that enhance the estimation of the *a*
586 *priori* statistics every time step in the analysis.

587 However, evaluating the uncertainty associated with RedAnDA remained challenging, encom-
588 passing various sources such as the quality and sampling of observations, the predictability of the
589 dynamics, and the unrepresented scales. Further investigation into each of these sources of error
590 will be studied in future works.

591 It is worth noting that assumptions regarding the validity of climatology and EOFs during the
592 reconstruction period, as they were calculated during the learning period may have had limitations.
593 As demonstrated in the OSSE, these limitations could have induced some errors, notably concerning
594 long-term changes in the climate system. Thus, estimating the associated uncertainty and under-
595 standing the relationship between temporal distance from the learning period and obsolescence are
596 crucial areas for future research.

597 In the OSSE, the non-stationarity of the mean state was found to be significant, particularly
598 at depth. Attempts to account for the changing climatology yielded unsatisfactory results. The
599 observations were insufficient in reflecting the long-term evolution adequately, likely due to uneven
600 sampling.

601 Space reduction constrained signal extrapolation at depth, where a greater number of EOFs were
602 required to fully capture variability. As previously stated, the curse of dimensionality imposed
603 limitations on the number of EOFs that could be utilized (?). Efforts were needed to enhance
604 the EOFs to encompass more of the variability present in the initial signal. Focusing solely on a
605 specific depth, it is feasible to adjust the relative significance of each level in the construction of
606 EOFs through weighting, allowing for the incorporation of the most crucial modes of variability
607 associated with the target depth.

608 The length of the catalog was constrained by the availability of Argo data, which might have
609 been insufficient to represent the full spectrum of ocean dynamics. Consequently, there was a risk
610 of underestimating the uncertainty in the analysis due to the potential redundancy of the selected
611 analogs and the inherent replication of associated trajectories. While using a catalog composed of
612 simulated data could have expanded its size, the primary objective was to generate an analysis solely
613 based on real observations. This approach aimed to avoid potential bias introduced by numerical
614 models, which could impact data assimilation despite observational constraints, particularly in
615 regions and periods with sparse sampling (??). Given that the influence of such factors was not
616 assessed in our case, it was decided not to incorporate simulated data into our reconstruction of
617 reality.

618 Within the OSSE framework, it was envisaged exploring a broader catalog by incorporating
619 additional members from the OCCIPUT ensemble (Penduff et al. 2014). However, since all 50
620 members were simulated by the same forced model with the same atmospheric forcing. In the case
621 of a forced OGCM, the variability associated with ENSO reduces to forced variability. Thus, it was
622 determined that the tropical Pacific basin, with its predominant wind forcing, might not provide a
623 significant enhancement in dynamics for the catalog (?). In regions where ocean internal variability
624 holds greater sway (e.g., the North Atlantic or the Southern Ocean), using multiple members may
625 be crucial for reliably assessing uncertainty.

626 For a real observational application, the catalog, climatology, and EOFs were derived from ISAS,
627 which already constituted an analysis of the Argo period. While the effectiveness of its optimal
628 interpolation had been demonstrated, its limitations – such as the lack of signal in shallow coastal
629 areas – undoubtedly constrained RedAnDA. In the OSSE, the learning period was optimal. One

630 method to evaluate the impact of the ISAS mapping technique would involve applying it to the
631 simulation prior to reconstructing the past and observing the resulting modifications.

632 The current analysis was intended for investigations into regional temperature interannual vari-
633 ability and at large time scales as well. Oceanic structures that were insufficiently observed during
634 the 20th century, notably at depth, may be characterized through the use of the new temperature
635 product provided by RedAnDA.

636 To reconstruct temperature and salinity with RedAnDA in a co-varying approach, could provide
637 new datasets of the evolution of regional water masses over the second half of the 20th century. To
638 consider their co-variability would be beneficial to salinity from the greater number of temperature
639 observations.

640 An illustrative application involves utilizing RedAnDA's Pacific temperature time series within a
641 new catalog to predict El Niño using analog forecasting. With the Precursors method (e.g., ???),
642 the analogs could provide new means of identifying conditions conducive to ENSO events.

643 *Acknowledgments.* Erwan Oulhen was funded by the Région Bretagne and the Université de
644 Bretagne Occidentale. This study was led in the framework of the LEFE GMMC OASIS project
645 (INSU/GMMC). ISAS temperature and salinity monthly gridded field products are made freely
646 available by SNO Argo France at LOPS Laboratory (supported by UBO/CNRS/Ifremer/IRD)
647 and IUEM Observatory (OSU IUEM/CNRS/INSU) at doi: <https://doi.org/10.17882/52367>. The
648 authors express gratitude to Thierry Penduff and Jean-Marc Molines for providing OCCIPUT
649 outputs and their invaluable assistance. Appreciation is also extended to Gaël Alory, Laurent
650 Bertino, Sally Close, William Llovel, and Jérôme Gourrion for their constructive discussions.

651 *Data availability statement.* ISAS20 data are available through [10.17882/52367](https://doi.org/10.17882/52367); EN4.2.2
652 profiles and objective analyses are available through [https://www.metoffice.gov.uk/
653 hadobs/en4/download-en4-2-2.html](https://www.metoffice.gov.uk/hadobs/en4/download-en4-2-2.html); OCCIPUT data are available upon demand; ORAS4
654 are available through [https://icdc.cen.uni-hamburg.de/thredds/dodsC/oras4_temp_
655 all.html](https://icdc.cen.uni-hamburg.de/thredds/dodsC/oras4_temp_all.html); IAP data are available through [http://www.ocean.iap.ac.cn/ftp/cheng/CZ16_
656 v3_IAP_Temperature_gridded_1month_netcdf/](http://www.ocean.iap.ac.cn/ftp/cheng/CZ16_v3_IAP_Temperature_gridded_1month_netcdf/); ISHII data are available through [https://
657 rda.ucar.edu/datasets/ds285.3/dataaccess/](https://rda.ucar.edu/datasets/ds285.3/dataaccess/); ESTOC are available through [https://
658 www.godac.jamstec.go.jp/jagdas/catalog/estoc/02c/catalog.html](https://www.godac.jamstec.go.jp/jagdas/catalog/estoc/02c/catalog.html). RedAnDA re-
659 sults and codes will be provided upon demand.

660 **References**

- 661 Baldwin, M. P., D. B. Stephenson, and I. T. Jolliffe, 2009: Spatial weighting and iterative projection
662 methods for eofs. *Journal of Climate*, **22** (2), 234 – 243, <https://doi.org/10.1175/2008JCLI2147>.
663 1.
- 664 Bessières, L., and Coauthors, 2017: Development of a probabilistic ocean modelling system based
665 on nemo 3.5: application at eddying resolution. *Geoscientific Model Development*, **10** (3),
666 1091–1106, <https://doi.org/10.5194/gmd-10-1091-2017>.
- 667 Boyer, T. P., and Coauthors, 2006: Noaa atlas nesdis 60. *World ocean database 2005*.
- 668 Bretherton, F. P., R. E. Davis, and C. Fandry, 1976: A technique for objective analysis and design
669 of oceanographic experiments applied to mode-73. *Deep Sea Research and Oceanographic*
670 *Abstracts*, **23** (7), 559–582, [https://doi.org/https://doi.org/10.1016/0011-7471\(76\)90001-2](https://doi.org/https://doi.org/10.1016/0011-7471(76)90001-2).
- 671 Cleveland, W. S., 1979: Robust locally weighted regression and smoothing scatterplots. *Journal*
672 *of the American Statistical Association*, **74** (368), 829–836, [https://doi.org/10.1080/01621459](https://doi.org/10.1080/01621459.1979.10481038).
673 1979.10481038, <https://www.tandfonline.com/doi/pdf/10.1080/01621459.1979.10481038>.
- 674 Dussin, R., B. Barnier, L. Brodeau, and J. M. Molines, 2016: Drakkar forcing set dfs5. *MyOcean*
675 *Report*.
- 676 Emery, W. J., and R. E. Thomson, 2001: Chapter 4 - the spatial analyses of data fields. 305–370,
677 <https://doi.org/https://doi.org/10.1016/B978-044450756-3/50005-8>.
- 678 Gaillard, F., T. Reynaud, V. Thierry, N. Kolodziejczyk, and K. von Schuckmann, 2016: In
679 situ–based reanalysis of the global ocean temperature and salinity with isas: Variability of
680 the heat content and steric height. *Journal of Climate*, **29** (4), 1305 – 1323, <https://doi.org/10.1175/JCLI-D-15-0028.1>.
681
- 682 Good, S. A., M. J. Martin, and N. A. Rayner, 2013: En4: Quality controlled ocean temperature
683 and salinity profiles and monthly objective analyses with uncertainty estimates. *Journal of*
684 *Geophysical Research: Oceans*, **118** (12), 6704–6716, [https://doi.org/https://doi.org/10.1002/](https://doi.org/https://doi.org/10.1002/2013JC009067)
685 [2013JC009067](https://doi.org/https://doi.org/10.1002/2013JC009067), <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2013JC009067>.

686 Hansen, J., M. Sato, P. Kharecha, and K. von Schuckmann, 2011: Earth's energy imbalance and im-
687 plications. *Atmos. Chem. Phys.*, **11**, 13 421–13 449, <https://doi.org/10.5194/acp-11-13421-2011>.

688 Kao, H.-Y., and J.-Y. Yu, 2009: Contrasting eastern-pacific and central-pacific types of enso.
689 *Journal of Climate*, **22** (3), 615 – 632, <https://doi.org/10.1175/2008JCLI2309.1>.

690 Kaplan, A., Y. Kushnir, M. Cane, and M. Blumenthal, 1997: Reduced space optimal analysis for
691 historical data sets: 136 years of atlantic sea surface temperatures. *J. Geophys. Res.*, **102860**,
692 835–27, <https://doi.org/10.1029/97JC01734>.

693 Kolodziejczyk, N., A. Prigent-Mazella, and F. Gaillard, 2021: Isas temperature and salinity gridded
694 fields. <https://doi.org/https://doi.org/10.17882/52367>.

695 Lguensat, R., P. Tandeo, P. Ailliot, M. PULIDO, and R. Fablet, 2017: The analog data assimilation.
696 *Monthly Weather Review*, **145** (10), 4093 – 4107, <https://doi.org/10.1175/MWR-D-16-0441.1>.

697 Penduff, T., B. Barnier, L. Terray, L. Bessières, and G. Sérazin, 2014: Ensembles of eddy ocean
698 simulations for climate.

699 Shea, D., 2013: National center for atmospheric research staff (eds). The Climate Data
700 Guide: Empirical Orthogonal Function (EOF) Analysis and Rotated EOF Analysis.
701 Retrieved from [urlhttps://climatedataguide.ucar.edu/climate-data-tools-and-analysis/empirical-orthogonal-function-eof-analysis-and-rotated-eof-analysis](https://climatedataguide.ucar.edu/climate-data-tools-and-analysis/empirical-orthogonal-function-eof-analysis-and-rotated-eof-analysis).

702

703 Sun, C., and Coauthors, 2010: The data management system for the global temperature and salinity
704 profile programme. *Proceedings of OceanObs*, **9**, 86.

705 Trenberth, K. E., and D. P. Stepaniak, 2001: Indices of el niño evolution. *Journal of Climate*,
706 **14** (8), 1697 – 1701, [https://doi.org/10.1175/1520-0442\(2001\)014<1697:LIOENO>2.0.CO;2](https://doi.org/10.1175/1520-0442(2001)014<1697:LIOENO>2.0.CO;2).

707 Zhen, Y., P. Tandeo, S. Leroux, S. Metref, T. Penduff, and J. L. Sommer, 2020: An adaptive
708 optimal interpolation based on analog forecasting: Application to ssh in the gulf of mexico.
709 *Journal of Atmospheric and Oceanic Technology*, **37** (9), 1697 – 1711, [https://doi.org/10.1175/](https://doi.org/10.1175/JTECH-D-20-0001.1)
710 [JTECH-D-20-0001.1](https://doi.org/10.1175/JTECH-D-20-0001.1).