Near surface ocean temperature uncertainty related to initial condition uncertainty

Robin Tokmakian¹ · Peter Challenor²

Received: 15 June 2017 / Accepted: 9 May 2019 / Published online: 16 July 2019
© The Author(s) 2019

Abstract

Contributions to climate uncertainty in the modeling of the near surface temperature of the ocean come from several sources: the knowledge of the initial state of the atmosphere, the inadequacy of a model to include or correctly represent all aspects of the dynamics, the uncertain knowledge of the initial state of the ocean and small-scale (or internal, “weather”) noise. This paper concentrates on the latter uncertainty, the uncertainty in the knowledge of the initial field, resulting in a map of the uncertainty of the near surface temperature field. To make the full estimate of this uncertainty, the outcomes from a relatively small, 30 member, ensemble were expanded using statistical emulators. The uncertainty varies in space and decreases in time. There are broad areas of similar uncertainty, consistent with uncertainties in the initial conditions and the variability in local processes. After 10 years, the average spatial uncertainty calculated from this expanded data set gives a range of ± 0.4 °C with 95% of the spatial areas having an uncertainty of less than ± 0.9 °C. When compared to other estimates of similar metrics, the map represents a lower bound on the uncertainty because this first experiment examines the ocean/ice system in isolation from feedbacks that occur between the atmosphere and the ocean. This uncertainty also does not include any contribution from the lack of knowledge about how a future state of the air/ocean system will be forced by changes in greenhouse gases.

Keywords GCM · Ocean · Emulators · Initial conditions · Surface Temperature · Uncertainty

1 Introduction

Substantial evidence shows that reasonable predictions of the climate can be made for periods up to 10 years (e.g. Smith et al. 2012; Branstator et al. 2012; Meehl et al. 2007) and such simulations were included in CMIP5 (Taylor et al. 2012). The detailed skill of such forecasts, on regional and smaller timescales, are uncertain and depend critically on the state of the climate, in particular, the ocean, used to initialize the forecast. As one of the components in our Earth system, a number of reasons supports the importance of the ocean in this context. The ocean retains the memory of the

air/ocean system because it is slower than the atmosphere to respond to changes at its boundary. If we want to understand the changes in the balance of heat between the ocean and atmosphere on scales of a decade or longer, how heat is absorbed by the ocean is important to predict correctly. Last, regional changes downstream on land at long timescales may also be influenced by the ocean’s state.

The question arises, “Can the uncertainty in a prediction on decadal scales be determined such that it will result in robust statistical distributions of a required metric and allow for risk assessment in successive analyses”. The uncertainty in a model includes uncertainty from different sources: initial conditions, the architecture of the model, the model parameter space, and the boundary condition forcing. This paper concentrates on the ocean initial condition uncertainty on decadal scales (5–15 years) using an ocean model forced at the surface with prescribed surface forcing fields. There is no feedback from the ocean back to the atmosphere. As such, this experiment lays the foundation for a following experiment which does include feedback processes. By doing this first experiment in an “ocean only” framework, it will allow

✉ Peter Challenor
P.G.Challenor@exeter.ac.uk

Robin Tokmakian
rtt8102@gmail.com

¹ Oc/tk, Department of Oceanography, Naval Postgraduate School, Bldg 232, Rm 328, Monterey, CA 93943, USA

² College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QE, UK

for the feedback effects to be quantified explicitly in a following experiment. A subsequent paper will describe those results when the feedback processes are included.

Smith et al. (2012) presented evidence on the importance of the initial conditions of the ocean for making reasonable forecasts on these time scales. They acknowledged that there are uncertainties in the initial conditions and are typically sampled by perturbing winds and ocean conditions or with a lagged average approach that combines forecasts starting from different dates. These methods of sampling the initial condition uncertainty are ad hoc and while they have provided a good starting pointing, they are not statistically robust. This paper presents a method that can robustly define the uncertainty. Using a set of reanalysis products combined with bootstrap methods, the probability space of initial condition uncertainties is formed. With careful sampling of the initial condition distribution space, a relatively small number of model simulations (or runs) of a forward, complex, dynamical, and deterministic model are executed and from the outcomes, an ensemble is created for a specific metric. The outcomes from the deterministic model are used to inform a posterior distribution of the metric in space and time by using statistical machinery to emulate relationships between the initial conditions and the metric (commonly referred to as emulators).

After reviewing some previous efforts in quantifying uncertainty for this particular problem, the experiment discussed in this paper is described. This involves both providing the details of the methods used for the creation of the initial condition space as well as the method for the determination of the posterior distributions (Sect. 4). Section 3 defines the experiment framework and with the application discussed in Sect. 5. Following the explanation of the posterior distribution and its related uncertainty quantity, the uncertainty quantity is compared to other, related metrics and their uncertainty Sect. 6). For the purposes of this paper, uncertainty is defined as $\pm 2\sigma$, where σ is the standard deviation of the quantity of interest. The near surface temperature field is our metric of interest because of its dominance as a metric in climate science.

2 Previous research in initial condition uncertainty

Smith et al. (2012) reports on the state of the science for predicting seasonal to decadal climate conditions. They discuss the importance of initial conditions on the long time scale forecasts as suggested by their Decadal Prediction System (Smith et al. 2007, DePreSys). This analysis illustrated the influence of initial conditions of both the ocean and land surface temperatures and the upper ocean heat content (0–300 m) for forecasts up to 9 years. Branstator et al.

(2012) examined six Atmosphere/Ocean General Circulation Models (AOGCMs) to assess the influence on predictions from the initial conditions. They found that anomalies in the ocean lasted approximately a decade for both the North Atlantic and the North Pacific. However, the size of the area influenced was highly variable between the different models, as high as a factor of 10. This was a synthetic analysis, in that no real observations were used to initialize the models. They determined that it was the mean of the forecast distribution, rather than the spread that was important in determining if a model included areas of predictability. Others (e.g. Matei et al. 2012) have found similar results relating to the importance of initial conditions. None of these papers quantifies the uncertainty as a function of the initial conditions.

Current efforts to quantify the uncertainty of a prediction at time scales from inter-annual to decadal are primarily based on small ensembles of outcomes that vary initial conditions and/or parameters of the ocean or atmosphere. The CMIP5 (Taylor et al. 2012, Coupled Model Intercomparison Project Phase five) program, for the first time, included a decadal prediction component. The protocol stated that these models be initialized every five years, starting in 1960. While the protocol indicates varying the initial point, there was no discussion of how to quantify initial condition error that may propagate forward. Having a start date every 5 years does vary the initial conditions. But, because the initialization has not been done in a formally designed experiment manner, that is, allowing the input state to be associated with an output state, an uncertainty quantification that is full and robust is difficult to achieve.

In previous experiments, the variations in the ocean initial conditions come from sampling a set of previous hindcast (with assimilation of data) runs (Pohlmann et al. 2013; Smith et al. 2007; Yeager et al. 2012; Troccoli and Palmer 2007) or as in Kröger et al. (2012) used initial conditions from three different reanalysis estimates. In some experiments (e.g. Yeager et al. 2012), all ensemble members are initialized with the same ocean, and it is the variation in the atmosphere and land components that uniquely determine an ensemble member. For example, Yeager et al. (2012) use 10 different days and/or differing hindcast runs for January of a specific year to initialize the atmospheric component for the 10 member ensemble, while Troccoli and Palmer (2007) use two different expressions of greenhouse gases (GHGs) for their six member ensemble. Most of these experiments differ from the one presented in this paper—in that it is the forcing at the ocean's surface that is responsible for the variation in the ensemble, rather than the initial conditions of the ocean itself. The one study (Kröger et al. 2012) that modified the ocean's initial conditions, only had three realizations of the outcomes. This limited ensemble is too small to be used to produce a robust uncertainty quantification or estimation.

As presented, these previous efforts relate to the framing of the uncertainty question as stated by Goddard et al. (2013). They pose two questions with respect to the initialization of decadal hindcasts: (1) "Do the initial conditions in the hindcasts lead to more accurate predictions of the climate compared to un-initialized projections?" and (2) "Is the prediction model's ensemble spread an appropriate representation of forecast uncertainty, on average?" This paper, on the other hand, addresses, a corollary question: "Given the uncertainty in our knowledge about initial conditions, how does this uncertainty affect a prediction?"

3 Experiment design

The experiment's goal, presented in this paper, is to understand the uncertainty in the near surface temperature field, given the uncertainty in the initial conditions. Consider an ensemble of simulations with some outcome, y , and where the outcome for each ensemble member, i , is defined as:

$$y_i = y_{\text{truth}} + \epsilon_{\text{initial},i} + \epsilon_{\text{numerical},i} \quad (1)$$

where y_i is the estimate of the metric given by one member of a forward model ensemble with some set of initial conditions, x_i . We assume that if a very large ensemble (many thousands) of simulations of the GCM could be conducted with different initial conditions, then the mean of those many simulations would converge on y_{truth} . The terms following y_{truth} are random variables representing different sources of uncertainty or errors: $\epsilon_{\text{initial},i}$ identifies the error related to an individual initial condition, i . It is the uncertainty, σ_{initial} , made up of all possible $\epsilon_{\text{initial},i}$, that is the focus of this paper. In addition, there is another error source needing to be addressed: $\epsilon_{\text{numerical}}$ is the error due to the numerical scheme (including the error due to missing dynamics) and is the same for each ensemble member because the same numerical code is used and the difference in ensemble members due to this error is zero.

When considering the uncertainty being addressed in this paper, σ_{initial} , each $\epsilon_{\text{initial},i}$ may have two contributions, $\epsilon_{\text{initial},i}^*$ and ϵ_{chaos} . The first contribution, $\epsilon_{\text{initial},i}^*$, is the uncertainty that is the same, every time the simulation is run with the same initial condition, i . ϵ_{chaos} represents a chaotic term, arising from the non-linear nature of the equations. That is, the dynamics of the model may take one or more paths due to the non-linear nature of the system (Schrier and Maas 1998), resulting in large differences in the solution after a period of time. Trying to find out if the system is chaotic after 10 years would be futile as we know from the Lorenz (1963) system that the effect is state dependent, i.e., for some initial conditions you would see it and for others you might not. A full investigation of the non-linearity of the model would involve the

estimation of Lyapunov exponents or something similar. This is a very complex task for a model of this size and is not done here. Two simulations with exact initial conditions were conducted and after ten years, the outcomes are indistinguishable. (This does not mean that after 100 or 1000 years, they won't diverge). Furthermore, (Hong et al. 2013) ran an experiment using the same GCM with the same computer architecture (Hong et al. 2013) and any differences were too small to measure. We address the chaos contribution in the discussion section.

A second mathematical framework describes the temporal variability. y_{internal} is the variability that is common to all model simulations, for example, the seasonal cycle and weather or short temporal frequency events (e.g. "weather"). In this paper, it is the temporal variation induced by various forces on a location. Given that each ensemble starts at the same time and has the same surface forcing, y_{internal} is the same. For any one ensemble member i , then

$$y_t = \frac{1}{T} \sum_{t=1}^T y_t + y_{t,\text{seasonal}} + y_{t,\text{trend}} + y_{t,\text{internal}}$$

and rearranging

$$y_{t,\text{internal}} = y_t - \frac{1}{T} \sum_{t=1}^T y_t - y_{t,\text{seasonal}} - y_{t,\text{trend}} \quad (2)$$

y_{internal} is what is described as internal variability (Sriver et al. 2015, e.g.). $y_{t,\text{seasonal}}$ and $y_{t,\text{trend}}$ are the portion of the signal due to seasonal forcing and any long term trend that is present. t is time and T is the total time period for the simulation, making the first term on the right side the long term mean for the simulation. In this framework, areas of similar patterns can be identified, but with each ensemble member having distinct phasing arising from the initial condition perturbation. This similar areas may help with identifying areas that can be used for prediction and those areas that can not be.

We use the first framework, Eq. 1, to determine, through the modification of the initial conditions, a spatial uncertainty map. This is achieved using wisely sampled initial conditions combined with an emulator to extend a limited ensemble of outputs. An emulator is a statistical model of relationships between the dynamical model and its initial conditions. An ensemble of 30 forward model simulations was created, each covering 10 years and each with a different initial temperature field. The ensemble size is limited because only 30 instances of the forward model can be computationally afforded. If there was sufficient computer power and time, the forward model could be run many thousands of times, each with a different initial condition

sampled randomly from the set of possible initial conditions. The outcomes are an ensemble that can be used to directly compute the probability distribution for the metric and thus, allow for understanding the uncertainty in a metric. The end result is to create a probability distribution for a deterministic model metric at a point in space and time, given that there is uncertainty in the initial condition. The experiment, as designed, does not support the ability to examine the outcomes, as related to inputs, based on scale selection (e.g. small spatial scale input uncertainty as related to similar spatial outcomes).

The second framework, Eq. 2, is used to help understand how important the initial condition uncertainty is relative to temporal variability within each model ensemble member.

The following paragraphs describe the forward model (or simulator) and the data from which the initial fields were derived using the methodology described in Sect. 4.1.

3.1 Forward model

The forward model or simulator being used in this experiment is the Community Earth System Model (CESM) (Collins et al. 2006). With a horizontal resolution of approximately 1° (110 km), the vertical structure is represented by 60 levels with the top 20 covering 5 to 200 m, approximately. The ocean component model is POP2 (Smith and Gent 2002), while the sea ice model is CICE (Hunke and Dukowicz 2003; Bitz and Lipscomb 1999). This experiment uses the model in an “ocean/ice only” mode, with the ocean surface forced by daily reanalyses winds and fluxes (Large and Yeager 2009, COREv2) for a 10 year time period covering 2000 through 2009. A 30 day relaxation term on the temperature field nudges the model towards the initial field over the first month. This is to avoid shocks to the system from the new initial condition.

Ten years of monthly averaged 3D prognostic fields for temperature, salinity, and velocity, along with 2D monthly fields of sea surface height are stored for analysis. Along with the prognostic variables, a large number of diagnostic quantities are available for analyses. The end result is an almost inexhaustible set of quantities which can be examined using the methods described in Sect. 4.2 to understand uncertainty of the model's processes due to differences in the initial conditions.

3.2 Initial field sources

Six ocean reanalyses contribute to the set of anomaly fields, the German Estimating the Circulation and Climate of the Ocean, GECCO (Köhl and Stammer 2008), the Geophysical Fluid Dynamics Laboratory—Ensemble Coupled Data Assimilation, GFDL-ECDA (Chang et al. 2013), the Predictive Ocean Atmosphere Model for Australia: POAMA

(Alves et al. 2003), the European Centre for Medium Weather Forecasting- operational ocean reanalysis system, ECMWF ORAS4 (Balmaseda et al. 2013), the University of Reading -analysis URA025.4 (Valdivieso et al. 2012) and the Simple Ocean Data Assimilation: SODA (Carton and Giese 2008). Anomalies were computed by differencing monthly fields from a reference reanalysis, URA025.4 for 2001 to the end of the particular reanalysis product (2009 or 2010). All the reanalyses use similar observational data sets but differ on how the interpolation of the data to non-sampled locations is performed. Very large scale signals, such the Atlantic Multi-Decadal Oscillation (AMO) or the El Niño Southern Oscillation (ENSO) patterns, are in the same phase for all reanalyses. Using the difference of a reanalysis from a reference field is a way to examine the differences in interpolations. It is a realistic method to sample what might be referred to as the plausible initial condition error.

For any location, i , the anomaly field (x^*) is defined as

$$x_{i,m}^* = x_{i,m} - x_{i,b}, \quad (3)$$

where m is the reanalysis product for January 2000, and b is the base field. URA025.4 is chosen as the reference field because it has the highest resolution. Any long-term trend in the difference and any residual seasonal signal (average January value for given m over length of reanalysis) were also removed. This process results in an ensemble of 564 anomaly fields from all the months available from all the reanalyses. Only potential temperature was used due to computational considerations. These fields are the input fields used in Sect. 4.1.

The base initial field has been extracted from one of the 20th century CESM large ensemble experiment data sets (Kay et al. 2015). For each of the 30 simulations, an anomaly field is added to the base field creating the total initial value field. The method for determining the anomaly fields is described in Sect. 4.1.

4 Input and outcome statistical methodologies

Two distinct statistical methodologies used in this study allow for the robust assessment of the uncertainty in the near surface temperature field due to the uncertainty in the knowledge of the initial field. The first method creates a distribution of initial value fields to sample from, while the second essential method builds a posterior distribution for some metric based on outcomes from a relatively small ensemble of forward simulations of the deterministic model. The posterior distribution is the basis for inferences about a metric's uncertainty. The following subsections describe the methods.

4.1 Input distribution and sampling

Initial conditions contain of two sources of uncertainty. First, is the uncertainty related to observational measurement error combined with any reanalysis error (in other words, how reanalysis method incorporates the observations). Uncertainty also varies with the state of the ocean. For example, in some cases, the ocean may be in a temporal state that is more stable on the short term (for example, a strong El Niño) verse a more unstable state (an eddy rich region). These regions may be spatially distinct. Thus, the inter-annual to decadal uncertainty or variance depends upon the physical state of the ocean and how intrinsically variable a region is This research focuses on the first, the uncertainty that is related to the initial condition given the state of the ocean, rather than the second. Our experiment is designed using a similar ocean initial state at large scales. This does not preclude the uncertainties introduced by the initial conditions migrate and may be enhanced, uniquely, by the similar surface forcing.

Because a complex climate model cannot be run the number of times required to produce a believable probability distribution, the set of all possible initial fields needs to be wisely subsampled. One possible solution might be to create a large number of initial fields (O500) and sample a small number (thirty) from the distribution in a non-random way to ensure adequate coverage of the set of initial conditions. However, the problem remains of how to create a valid distribution for the outcomes, given only thirty members of some spatial field. It would be difficult, if not impossible, to find a regression relationship between the initial value spatial field and an outcome that could be used to statistically expand the set of possible outcomes. Such a regression is necessary to form a convincing distribution.

However, by reducing the initial condition space substantially, such a relationship can be determined. The 4-dimensional space (3 spatial locations: latitude, longitude, depth and one sample number) decomposes, using a standard method, into a set of principal components (PC) (Pearson 1901; Hotelling 1933). This method allows for the reduction of its dimensions in the form of a set of spatial maps

and associated loadings or amplitudes. The PC loadings are sampled and used with the associated PC spatial maps to rebuild an initial value spatial field which, in turn, is used to initialize the forward model. The sampled loadings can be treated as the independent variables (rather than the fields themselves) for a multi-variate regression process (the emulator, see Sect. 4.2) to relate the initial value fields to a set of outcomes from the forward deterministic simulations.

The next step is to sample the loadings to achieve a realistic representation. Figure 1 shows a sample histogram for a set of values with a variance of 1.1, that will be used to illustrate different sampling strategies. The loadings could be sampled, for example, using a Gaussian or uniform distribution. The top two lines of black stars represent the locations in the distribution of the loadings sampled if a Gaussian strategy is used for two different trials. It results in variances for the sampled points of 1.2 and 1.3 and only in one case is an extreme value of the original distribution sampled. Employing a uniform strategy (bottom two sets of black dots) results in variances from the sampled points of 3.3 and 3.8, much higher than the original distribution. But if a sparse space filling design such as a Latin Hypercube (McKay et al. 1979), which is used in this study, is used to sample the loadings, the extremes as well as the interior of the multi-dimensional input space (the loadings) are sampled. The example sampling, represented by the plus sign symbols, samples the extremes and has a variance similar to the original distribution.

Because of computational limitations, our forward model simulations are limited to an ensemble size of 30 members. Faced with the problem of how to sample our initial field ensemble in such a way that it will allow for the interpolation of a much larger synthetic ensemble, size 5000 or more, from the outcomes of the 30 simulations, the following steps were taken. The first step decomposed the data set described in the previous paragraph into PC loadings and spatial maps. The size of the data set (described in (Sect. 3.2)) is large (76641 spatial points by 30 levels by 564 samples) and it was necessary to divide the data set into regions before the decomposition. The regions were defined as the ocean

Fig. 1 Idealized sampling strategy. Histogram with a variance of 1.1. Each line of 30 symbols represent different sampling strategies. Black stars represent uniform sampling and the black dots represent Gaussian sampling, plus signs represent Latin Hypercube strategy. Corresponding variances of sampled points given to the right of the symbols

basins: (Indian, Pacific, Atlantic and Southern Ocean). Once decomposed, 30 samples were chosen from the 564 samples for each of 120 PCs using a Latin Hypercube scheme as described above. Once the loadings had been identified, they were combined with the appropriate spatial field to create a recomposed initial field anomaly field. These anomaly fields were added to the base field (Sect. 3.2).

Figure 2a illustrates the distribution of the loadings for the first mode, a fairly Gaussian distribution (shown by the solid black line). The open circles locate the sampled loadings for one mode, that is 30 sampled loadings. (Figure 2b is the histogram of those values). Figure 2c illustrates the sampled loadings for the one initial condition from across

the PC spatial maps (sample number). The grey shading are the values of the loadings for that particular spatial map or sample. Thus, the initial condition field, when reconstituted, captures both the extremes across the spatial extent and also across the width of individual mode.

Figure 3 shows an example of one of the initial value anomaly temperature fields, reconstituted from the PCs. It contains large patches of similar sized anomalies as well as areas that have changes on much smaller spatial scales. The mid-latitudes anomalies are of smaller scale with larger amplitudes than what is seen in the tropics because of the nature of the differences in the reanalyses (source fields). These are, generally, in the regions of the observed ocean that are highly energetic with much eddy activity.

Fig. 2 **a** Histogram of loadings for 500 samples. Symbols above represent location of sampled loadings for one mode. **b** Histogram of PC loadings for one initial condition. **c** Typical sampling for on initial condition

Fig. 3 Example map of an anomaly field for one initial condition

4.2 Posterior or outcome distribution

After creating the forward model ensemble with different initial conditions, sampled outcomes (where the outcome is some metric of interest) define the set of the dependent variable for the uncertainty quantification. The posterior distribution is found for these outcomes given the uncertainty in the initial conditions through the use of the machinery of a statistical emulator. This emulator allows for the determination of an outcome at locations where the simulator has not been run. Thus, through the use of the emulator, an ensemble for 5000 inputs can be created. The brief summary of the method provides a reader unfamiliar with the concept the necessary background. Full details are found in appendix A.

In its simplest form (2 dimensions), the emulator is a regression fit of the input data (in this case, the loadings used to build initial conditions) to some output metric. Based on Gaussian process (GP) regression, an emulator has the advantage that it is more flexible than linear regression methods (and handles non-linear relationships) and is as adaptable as neural networks but easier to interpret. GP methods also give estimates of the emulator uncertainty. A good, general reference for GPs is Rasmussen and Williams (2006). Challenor et al. (2010), Urban and Fricker (2010), Rougier and Sexton (2007), Holden and Edwards (2010) provide examples of GPs for use with complex numerical models of oceanic and atmospheric processes. Previously, the authors used emulators based on GP to understand the parameter space of imbedded physical models within geophysical deterministic models (Tokmakian et al. 2012; Tokmakian and Challenor 2014).

Statistical inference about a quantity of interest for the simulator is carried out not using the simulator but using the emulator.

5 Application results

Thirty simulations of the forward model have been run with varying initial conditions. This is followed by the creation of an emulator for the outcomes of interest so that an ensemble of outcomes of sufficient size can be generated to allow uncertainty to be calculated. Figure 4 illustrates the sequence of events that lead to the creation of robust uncertainty estimation. The events are described in the following paragraphs.

5.1 Forward model outcomes and the metric of interest

The outcomes at the end of the 10 years are extracted from the simulation model fields. For this particular problem, the deterministic model ensemble members are sampled at the

Fig. 4 Schematic describing the sequence of events undertaken in this study. The gray boxes and circle indicate the elements related to the emulator creation. The letters in parentheses are used in the text for identification

10 year mark with the understanding that 10 years is a climate time scale rather than a weather temporal scale. The metric considered in this paper is the spatial temperature at about 25 m depth, or level 3 of the model. (The surface temperature is not used because of the large influence from air/sea interactions, given that the forcing data set is the same for each ensemble member). Once all the simulations are completed, the output is sampled for use with the emulator machinery. For most of the analysis in the paper, the output of near surface temperature (the metric of interest) is averaged to 5° by 5° squares along with the removal of the mean of the 30 member ensemble. This is done to reduce the number of emulators need to be created (one for each spatial point) as well as to reduce the shorter spatial scale variability. The shorter spatial scale variability (1° boxes) is examined in detail for the North Atlantic region. The end result is, for each square, a set of 30 temperature anomaly values that provide the outcome (y, Eq. 1 for the generation of an emulator (Fig. 4a).

5.2 Emulators of near surface temperature

To create the uncertainty distribution, the inputs and outputs are first defined. The outcomes are the 25m near surface temperature anomaly in 5° boxes and are associated with the inputs which are the PC loadings used to create the initial value fields (gray boxes, Fig. 4a, b). An emulator (see Sect. 4.2) is then created for each 5° by 5° ocean grid box for the temperature at level 3 (25 meters) of the model.

An emulator is created for a 5° box centered at 42.5°N , 292.5°E in the North Atlantic, as an example. It is validated

through the use of a “leave one out” strategy. That is, the “regression” or emulator code uses only 29 of the 30 outcomes to estimate the regression coefficients. The 30th outcome tests the robustness of the regression using the regression coefficients from the other 29. This exercise is repeated 30 times, each time leaving out a different member of the ensemble. If 90% of the 30 tests (27) are successful, then a similar emulator using all 30 outcomes is assumed to be reasonable also.

For the example, Fig. 5 shows the 30 forward model values verse the mean estimate for that outcome from the emulator. These are the “left out” values for each of the 30 emulators. The bars are the extent of the uncertainty (2σ) associated with each of the mean values. The emulator gives a mean value plus an estimate of its uncertainty in that value. 29 of the 30 estimated mean values are within range of the forward model’s value, or 97%, satisfying a criterion of better than a 5% significance level. Having a validated emulator allows for the creation of a full uncertainty distribution. By using the emulator to estimate outcomes for a set of initial fields (inputs), the emulator gives an estimate of an output for an initial field not tested with the dynamic model (due to a lack of computer power to test all locations).

Using the validated emulator, a set of 5000 outputs are generated. This results in a posterior distribution (see 4.2) for the near surface temperature, shown in Fig. 6a, right histogram. The values are the emulator’s mean estimate of the metric at 5000 input locations (Fig. 6a, left histogram). Along with the mean, a variance is also obtained for each location (Fig. 6a, middle histogram). For this location, the set of possible mean temperature anomaly values range between about ± 0.5 °C. The distribution of the

uncertainty (or variance) about these mean values has a maximum of 0.15 °C. Combining the information for the mean distribution and the uncertainty about the mean gives a complete distribution of possible temperature anomalies for the location, shown in Fig. 6a, right histogram. Therefore, at the location of 42.5°N, 292.5°E, given a set of varying initial conditions, after 10 years, the possible temperature may vary about ± 1.1 °C, with 95% of the values between -0.52 °C and 0.49 °C. The uncertainty arises both from the initial anomalies introduced within the analyzed box and also from anomalies propagating into the area. It is not possible to separate these two aspects, but both are associated with the initial condition anomalies. If it were a true Gaussian distribution, the 95% location would be ± 0.49 °C. Thus the distribution is approximately Gaussian (the black line in Fig. 6a, right histogram overlay). Further examples of the three histograms are shown for a tropical region, Fig. 6b and a polar region, Fig. 6c. The locations are indicated on the Fig. 7’s map with a white X.

5.3 Uncertainty of the near surface temperature field

Section 5.2 illustrates how uncertainty is computed for one location. Figure 7 shows this computation of uncertainty for every 5° square location for the ocean domain of 0–360°E, -90 – 75 °N. The map displays the 2σ value of the uncertainty computed by the emulator (see Fig. 6, right column). Plus and minus the mapped value gives the range of the uncertainty for a given point. PDFs (Fig. 8) of the uncertainty at each location, overlaid one upon the other, illustrate the breadth of the distributions globally. There are only a few areas (less than 1%) with uncertainty greater than ± 2 °C, in the North Atlantic and in the North Pacific around 45°N, 165°E. Large patches of uncertainty in the range of ± 0.5 °C are spread throughout the oceans. The average uncertainty is about ± 0.4 °C with 95% of the locations having an uncertainty less than ± 0.9 °C. These cohesive regions illustrate that the uncertainty is not just random noise. Rather, it is related to the physical processes occurring in the region and how the observational sampling of the ocean within the regions contributes to the uncertainty in the initial conditions. For example, in the North Pacific and North Atlantic in the region of the western boundary currents and their extensions, eddy activity results in the knowledge in initial condition is less certain. These areas with uncertainty greater than ± 0.5 °C are known to be more energetic. The low uncertainty areas can result from two situations, both a result of having no feedback between the ocean and atmosphere in this ensemble. As time progresses, the initial conditions are less important and the similar atmospheric forcing is the dominant contributor to reducing the uncertainty with time. There are differences,

Fig. 5 Validation of emulator at one location in the North Atlantic, 42.5°N, 297.5°E. Bars represent the uncertainty at the location for the point left out in the validation test

Fig. 6 **a** Histograms of mean emulated near surface temperature values given an initial condition distribution for the location 42.5°N, 297.5°E in North Atlantic, variances associated with each mean value

from the emulator for the location and the mean + 2σ uncertainty of the emulator. **b** The same 3 histograms for a tropical location, 0°N, 160°E; and (c) for a location in the polar region, 60°S, 170°E

Fig. 7 Map of maximum positive uncertainty. Color scale is logged. Color legend labels are in $^{\circ}\text{C}$. The white Xs mark the locations of Figs. 5 and 6 analyses

Fig. 8 PDFs overlaid for all locations illustrating breadth of uncertainty range

even after ten years. If the inputs (e.g. the tropical Pacific region, see Fig. 3) have low uncertainty (as compared to a more active region such as the North Atlantic Current area) due to a high number of observations, then the forcing will dominate and the ensemble members converge as time progresses. Given the long term measurements from the TOGA/TOA (<https://www.pmel.noaa.gov/tao>) array in the Pacific, this low uncertainty is not surprising. Secondly, if there is a lack of measurements (area south of 70°S) contributing to the initial conditions, reanalyses will be similarly interpolated. And, again, the result is that there will be low uncertainty in the outcomes.

A temperature change in the ocean can be reflected as a change in steric height, $h_s = \int \frac{\rho_0 - \rho}{\rho_0} dz$ where ρ_0 is a reference density and ρ is the density of the water at new temperature. In some cases, it may be preferable. The temperature uncertainty translates to about a maximum of ± 5 mm for a 1 °C change over a 25m layer of surface water, without considering the change in the waters below the surface layers.

The temporal change in the uncertainty of the outcomes induced by the uncertainty in the initial field is illustrated in Fig. 9a. Shown as a zonal average, the uncertainty in the model's solution is reduced over time, implying that the differences in the initial fields' (black line) influence diminish, on average, with time and the solutions converge. In other words, the initial conditions have less and less impact on the solution as the simulator progresses in time. This is expected because the near surface temperature field is strongly influenced by the atmospheric forcing. In some regions, for example, the tropics, the uncertainty is greater than the year before (2007 verse 2006, $y_{t, internal}$, Eq. 2) but much less than the initial uncertainty. If these maps of uncertainty were to be used in an external application, a

better approach might be to compute an uncertainty map using years 5 through 10. In a coupled system, this inter-annual uncertainty is included in the uncertainty at the end of a 10 year simulation because of the influences of feedback processes. Note, again, that there is no feedback from the ocean back to the atmosphere in these simulations. (This is the subject of a following paper.)

5.4 Flow scale dependencies

To understand the influence of the smaller scale processes (at 1°, the resolution of the simulator) on uncertainty, maps of uncertainty at this resolution for the region of the North Atlantic were created. The uncertainty maps were created by first averaging the temperature fields from the model, followed by performing the uncertainty analysis using these fields at a specified resolution. Four years (3, 6, 8, and 10) were examined at three resolutions (1°, 5°, and 10°) plus one addition map that considered a 1° resolution field with large scale features removed. This quantity, $y_{i,j}^*$, is defined mathematically, for every location, i, j as:

$$y_{i,j}^* = y_{i,j} - \frac{\sum_{i=2}^{i+2} \sum_{j=2}^{j+2} y_{i,j}}{n}, \quad (4)$$

where n is the number of non-land locations being summed over an approximately 5° box.

At each of these resolutions and at every location, an uncertainty calculation was performed for December of year 3, 6, 8 and 10. The results are shown in Fig. 10. Each column represents a different resolution: 1° minus 5° is the first column, followed by the full 1°, 5°, and 10° fields. Each row represents a different year, with year three at the top and year 10 at the bottom. The number in parentheses in the title of each subplot is the 2σ average value for that plot.

As expected, all columns show a decrease in uncertainty from year 3 to year 8. In year 10, at the smallest resolution, there is an increase of about 9% in variance, due to small scale features. In other words, because the overall average between Fig. 10k, o is the same, the difference in Fig. 10j, n comes from the difference between Fig. 10i, m, the small scale uncertainty. Furthermore, between the years 8 and 10, the uncertainty has reached a stable state at the larger scales. There is no evidence that the difference in the small scale is anything but internal variability (Eq. 2. Table 1, column 3 shows that for years 8 and 10 the larger scale (5° and greater) processes are contributing about half to the uncertainty when averaged at (1°) (Table 1, column 2). The other half is the contribution for the processes that

Fig. 9 **a** Comparison of zonal average of near surface temperature (25 m) uncertainty for December of model simulation (color lines). Gray line indicates the uncertainty in comparison to a control run's

December variance. Black line indicates uncertainty in the initial conditions. **b** Ratio of initial condition uncertainty to control run uncertainty

on the order of 1° (Table 1, column 1). This division of uncertainty by spatial scales gives an approximate indication of how various process at different scales influence the uncertainty.

Years 8 (2007, as determined by the forcing fields) and 10 (2009) of the simulation represent two different states of the North Atlantic Oscillation (NAO) with index values of 1.34 and -1.93 , respectively (Barnston and Livezey 1987, see NAO index at Climate Prediction Center). These two years can be used to examine how interactions

between the large scale and smaller scales may be influencing the uncertainty maps. As Fig. 10i, p show (having the same average uncertainty range of $\pm 0.23^\circ\text{C}$), there is little difference in the largest scale uncertainty pattern. The differences in uncertainty come from smaller scales anomalies (see Fig. 10i, m) in the outcomes. While this experiment's outcomes are conditional upon the state of the system in year 2000, the experiment produced no evidence that shows that interactions between the large scale ocean state and the smaller scale differences between

Fig. 10 Maps of North Atlantic uncertainty (2σ , $^{\circ}\text{C}$) for 4 resolutions and 4 years. Column 1: 1°-5°; Column 2: 1°; Column 3: 5°; Column 4: 10°. Row 1: December of year 3; Row 2: December of year 6; Row

3: December of year 8; Row 4: December of year 10. Color range is for a logged value between -3 and 1 (corresponding to an uncertainty range of 0.001–10 $^{\circ}\text{C}$). In parentheses is the average 2σ value for the map

Table 1 Variance (σ^2) values for Fig. 10, $^{\circ}\text{C}^2$

Year	1°-5°	1°	5°	10°
3	0.06	0.31	0.18	0.10
6	0.04	0.11	0.04	0.02
8	0.02	0.06	0.03	0.01
10	0.03	0.07	0.03	0.01

ensemble members influence the resulting uncertainty maps.

A similar analysis has been performed for the tropical Pacific region (Fig. 11) for model years 7–10 (corresponding to actual years 2006–2009). The uncertainties are half as large or smaller than those in the North Atlantic at all resolutions. The December NINO3.4 anomaly index values (Trenberth 1997, see NCAR climate index site) for years 2006–2009 (simulation years 7–10) are 1.09, -1.24, -0.75, 1.43, respectively. 2007 and 2008 (year 8) are La Niña years, 2006 and 2009 are El Niña years. The two El Niño years (Fig. 11a–d, m–p) do not exhibit any large scale influence of

the large scale on the smaller scales. Only in year 8 (2007) is there a hint (Fig. 11h) that the large scale is influencing the smaller scales. However, it is quite small when compared to the uncertainties in either the very small scale map (Fig. 11e) or the 5° average map (Fig. 11g). Because the atmospheric forcing is strong in the tropics and is the exactly the same for all simulations, one might expect uncertainties in the near temperature field to be similar for all ensemble members.

6 Discussion

The significance of the mapped near surface temperature uncertainty as a function of initial temperature uncertainty can be assessed by comparing the magnitudes of the uncertainty to four other uncertainty and variance quantities: (1) internal variability of temperature over one year within one simulation (Eq. 2), (2) uncertainty in initial conditions, (3) uncertainty of temperature at 100 m, and (4) uncertainty of

Fig. 11 Maps of tropical Pacific (-10 to $+10$ latitude, -160 to $+160$ longitude) uncertainty (2σ , $^{\circ}\text{C}$) for 4 resolutions and 4 years. Column 1: $1^{\circ}\text{--}5^{\circ}$; Column 2: 1° ; Column 3: 5° ; Column 4: 10° . Row 1: December of year 7; Row 2: December of year 8; Row 3:

December of year 9; Row 4: December of year 10. Color range is for a logged value between -3 and 1 (corresponding to an uncertainty range of $0.001\text{--}10$ $^{\circ}\text{C}$). In parentheses is the average 2σ value for the map

temperatures from a process parameter variation ensemble. Each comparison places this study in context to other measures of variance or uncertainty.

6.1 Internal variability comparison

The first comparison assesses the uncertainty in the knowledge of the surface temperature value as compared to the intrinsic variability of the ocean over a year due to surface forcing and physical processes occurring within the ocean. The variance of the last year of one ensemble member is calculated after removing a monthly mean from the last five years of the simulation. The removal allows the variance to be calculated without a seasonal signal. This removes part of the influence of atmospheric conditions. By using only one year to estimate the variance, any inter-annual signal (strongly coupled to atmospheric influence) is ignored. What is left, for the most part, is the internal variability of

the ocean in the model. For comparison purposes, the variance is represented as twice the standard deviation. Shown in Fig. 9a as the gray line, the zonal average of the internal variability ranges between approximately 0.5 $^{\circ}\text{C}$ and about 1 $^{\circ}\text{C}$. It is less or equal to the ensemble uncertainty values for the latitudes greater than about 20°N for the first four years of the simulation period (years 2001–2004, dotted lines). At latitudes greater than about 30°N , in the latter years, much of the internal variance is about the same order as the uncertainty introduced by varying the initial conditions (Fig. 9b). In the tropics and southern hemisphere, the ensemble uncertainty due to the initial conditions is about half or less than that of the internal variability, especially in the last half of the simulation time period. Thus, over large areas, the initial condition ensemble uncertainty is fifty percent or greater than the non-seasonal, non-inter-annual internal variability. In contrast, the tropical variability is primarily due to internal variability rather than the uncertainty

introduced by varying initial conditions. The scale separation analysis, shown in Fig. 11, confirms this. Using the year with the largest uncertainty (year 8), the small scales contribute approximately two thirds of the variance to the large scale (variance ratio of 1° to 5° averages, 2σ average map value divided by 2 and squared, or 0.004/0.006).

This analysis illustrates that the uncertainty introduced by different initial conditions, after 10 years, may still be influencing the model's solution 10–100%, depending upon location. This does not say that an initial condition perturbation of smaller or larger magnitude would result in the same outcomes. An additional experiment would be necessary to make such a conclusion.

6.2 Uncertainty at 100 m comparison

The previous sections (Sect. 5.3) considered the uncertainty in the temperature field at a depth of 25 m. At this depth, there is considerable opportunity for these waters to interact with the ocean's surface and, thus, be influenced by the atmospheric properties through mixing properties. Similarly, the surface waters are strongly influenced by the atmospheric forcing conditions, especially since this model configuration has no feedback from the ocean to the atmosphere. Examining the uncertainty at a deeper level (100 m) (Fig. 12), shows that the uncertainty drops after the first few years to a level that is seen in the zonal average at 25 m.

6.3 Comparison to parameter uncertainty

In a previous study on the influence of model parameter settings (Tokmakian and Challenor 2014), using a similar ocean model (POP2), but at a different resolution (3°), significant uncertainty in the model's solution was found to relate to the a set of process (e.g. mixing, advection) parameter values. The uncertainty in the upper ocean heat content is driven mostly by parameters related to the mixing scheme of the model, rather than parameters related to horizontal advection. For a relative comparison of the current study to this previous research, an analysis of the temperature field at about 25 m after 100 model years using 80 simulations was conducted. Following the same procedure as used for the initial field ensemble, uncertainty values for each 5° averaged area were computed for the parameter ensemble. The uncertainties for the parameter setting runs (Fig. 13b) are an order of magnitude greater than that of the initial field ensemble (Fig. 13a). The zonal structure of the two ensembles is similar; the higher latitudes have greater uncertainty than the tropical regions. However, the east-west symmetry differs. In the parameter ensemble, the uncertainty in the North Pacific is much lower than in the North Atlantic, while in the initial field ensemble, the tropical Pacific's uncertainty is lower than in the tropical Atlantic. Conversely, the initial

Fig. 12 Comparison of zonal average at depth of 100 m, uncertainty for December of year of model simulation (gray lines). Black line indicates uncertainty in initial conditions

field ensemble shows a similar level of uncertainty in the North Pacific and North Atlantic regions while in the tropical areas of the parameter ensemble, the pattern of uncertainty shows consistency across the three basins (Atlantic, Pacific, and Indian).

6.4 Comparison to other initial condition uncertainty estimates

It is difficult to make a direct comparison of this study with others because of the differences in experiment design and analysis. Kröger et al. (2012) is the closest to this study in how the initial conditions were prescribed.

Fig. 13 Spatial comparison of uncertainty maps between near surface temperature uncertainty conditioned on (a) initial condition uncertainty (after 10 years) and (b) parameter uncertainty (end of 100 years). Note the parameter uncertainty maps are formed from 3°

average grid cells rather than 1°. The color scales are different by an order of magnitude. Color scale is logged. Color legend labels are in °C

Similar to this study, Kröger et al. (2012) also used reanalyses to vary their initial conditions. But it is a coupled simulation study, rather than an ocean-only study. Their study examined three simulations which each used a different initial condition. Given these caveats, their Fig. 1a shows an anomaly range of temperatures at the surface averaged over the area 80–0°W, 0–60°N of between 0.4 and 0.55 °C at the end of their simulations. This is a similar order of magnitude as our zonal averages. Their Fig. 7, gives a spread of standard deviation values for the North Atlantic area of between 0.2 °C and 0.25 °C, or a difference of 0.05 °C, which if converted to the definition of uncertainty in this paper (2σ), is 0.3 °C. Again, this is on the same order as the uncertainty that is found for this study. If it is hypothesized that a coupled simulation should have more uncertainty than an uncoupled simulation (this study), then the Kröger et al. (2012) small ensemble underestimates the uncertainty induced by different initial conditions. A conclusive answer will be provided in the follow on paper which includes feedback processes.

In the Smith et al. (2007) study of a coupled climate system (initialized by varying atmospheric and oceanic conditions in four integrations), the stated range of uncertainty for 9 years in the future was on the order of ± 0.35 °C, based on their Fig. 2b for the full initial field or ± 0.2 °C when only an anomaly is added to the initial field (their Fig. 2c). This is a global average, rather than a zonal average. The first implies a much larger uncertainty than is found for the uncertainty of initial conditions of the ocean alone without feedbacks and the second is on the same order as this study. With only four ensemble members in the study, the uncertainty is arguable as to its value. However, it does provide a benchmark to evaluate the uncertainty estimates stated in this paper.

6.5 Initial condition uncertainty and chaotic uncertainty

There remains an issue about how much of the uncertainty might be due to possible chaotic nature of the flow. We took two simulations that had the smallest root mean square difference (RMSD) in their initial conditions and found the spatial temperature RMSD between the outcomes at the end of years 1 through 10. The RMSD values fell consistently from 0.17 to 0.04 °C. This implies that the two simulations are converging. If a chaotic term is influencing the dynamical system, the solutions would be diverging (Lorenz 1963; Li and Yorke 1975). Furthermore, if there were significant chaotic variability, the emulator, being a smooth function, would not be able to predict the uncertainty related to chaos. Using the leave one out validation (Fig. 5) we can compare the variation (variance) across the ensemble that is explained by the emulator compared to the total variation. The percentage explained by the emulator is 88%. The other 12% is made up of small scale variability not captured by our emulator plus any variability arising from chaos in the model. In our opinion this shows that the uncertainty coming from any chaotic variability is small compared to the initial condition uncertainty.

Taking into account the evidence just presented, we leave it to readers to determine if this dynamical system's variability, after 10 years, is due to chaos or the initial conditions. Because all members of the ensemble analyzed in this study begin at the same temporal starting location and all members have the same numerical code, we assume that the error sources, $\epsilon_{\text{numerical}}$ and ϵ_{chaos} , are, either the same or too small to measure and the only contribution to σ_{initial} is the initial condition error, $\epsilon_{\text{initial}}^*$.

7 Conclusions

Maps of near surface temperature uncertainty, after 10 years of integration, show the influence of the uncertainty in ocean initial conditions on the spatial temperature field. The maps indicate a higher uncertainty in the regions outside the tropics, areas generally expected to be more energetic. Physically, the perturbations introduced in the initial conditions results in uncertainties that are larger than what is seen in the internal variability for regions that are, temporally, more active at the smaller scales. In regions where the dominant temporal variability is at large spatial scales (e.g. the tropics), the initial perturbations dissipate because of the dominance of the surface forcing. This conclusion is supported by the analysis summarized in Figs. 9, 10, and 11, Sect. 5.4, in part.

While this study does not include the feedback processes that are found in a coupled air/ocean climate models, the framework sets the foundation for the methods to be used when examining a coupled system. With more and more interest in the use of climate models and their output for subsequent risk analysis on small or regional scales, these uncertainty maps bound the lower end of the uncertainty range. Even without feedbacks and ignoring seasonal and inter-annual variability due to atmospheric processes, the uncertainty of the near surface temperature field due to varying initial conditions after ten years ranges from 0.01° to 1°C , depending upon the region (translated to about a maximum of ± 5 mm). This uncertainty should only be used as a lower bound in any risk analysis because feedbacks have not been included.

Acknowledgements This work was completed with funding under NSF Grant No. 0851065. We thank the contributions from the CESM Large Ensemble Group and the various ocean reanalyses projects: GECCO, GFDL-ECDA, POAMA, ECMWF ORAS4, URA025.4, and SODA. The National Center for Atmospheric Research contributed the computational resources necessary to complete this research. We thank the reviewers for their constructive comments.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix A Gaussian process emulator details

Bayesian analysis, at its essence, combines a prior distribution (informative or not) with the likelihood of a data set to produce a conditioned posterior distribution for the data. Formally

$$\pi(\theta|x) = \frac{\pi(x|\theta)\pi(\theta)}{\pi(x)}, \quad (5)$$

where the term on the left hand side is the posterior pdf of the metric of interest (denoted θ) given the x , the data (in this case, the initial field loadings). This is equal to the likelihood of observing the data, x , given the metric (θ) multiplied by the prior pdf of θ . It is divided by the prior pdf for the data. This is a normalizing constant and can be estimated because the posterior pdf has to integrate to one.

Using the Bayesian framework, a prior defines the form of the GP. The general form of the GP for the prior mean function is given by: $m_0 = h(\bar{x})^T \beta$, where $h(\bar{x})^T$ is a vector of q regression functions and β is a vector of q parameters. For this study the mean prior function is a linear function with a zero mean. This means the joint distribution of any two points is Gaussian with the covariance given by $v_0(x_1, x_2) = \sigma^2 \chi(x_1, x_2)$ where χ is a correlation function.

The posterior mean function is not equal to the prior mean function. Rather, it combines the prior covariance function and the data. The formal expression for the posterior mean is:

$$m^* = h(\bar{x})^T \hat{\beta} + t(\bar{x})^T A^{-1} (Y - H \hat{\beta}), \quad (6)$$

where $\hat{\beta} = (H^T A^{-1} H)^{-1} H^T A^{-1} Y$. A is a n by n covariance matrix of the data (outcomes of forward model) with itself and t is the n by 1 covariance matrix between the data and any new value x . In this case, a Matern covariance function. The general form is

$$\chi(x_1, x_2) = \prod_{i=1}^q \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}(x_{1,i} - x_{2,i})}{b_{ii}^2} \right)^\nu \times K_\nu \left(\frac{\sqrt{2\nu}(x_{1,i} - x_{2,i})}{b_{ii}^2} \right), \quad (7)$$

where $x_{j,i}$ is the i th parameter for a given location j , b_{ii} is the smoothing parameter in that dimension, q is the number of parameters, and K_ν is a modified Bessel function (with its arguments following in the brackets) and Γ is the Gamma function. The emulators in this paper use $\nu = 7/2$. H is the matrix of the prior mean function. The first term on the right hand side of Eq. 6, determined from the prior mean with respect to the outcomes, is a regression function. The relationships between the different members of the model response, Z , and the initial conditions, x of the second term modify the regression function. As used for emulating complex numerical models and because the simulator is a deterministic model, the outcomes from the simulator are exactly interpolated. A “nugget” term, reflecting a misfit between the data and the GP at small scales is included. Away from the data points, the second term converges to zero and the fitted model reverts to the form of the prior.

A posterior covariance term that gives us the uncertainty of the predictions is defined as:

$$\begin{aligned} v^*(x_1, x_2) = & \hat{\sigma}^2 [\chi(x_1, x_2) - t(x_1)^T A^{-1} t(x_2) \\ & + (h(x_1)^T - t(x_1)^T A^{-1} H) \\ & * (H^T A^{-1} H)^{-1} (h(x_2)^T - t(x_2)^T A^{-1} H)^T], \end{aligned} \quad (8)$$

where $\hat{\sigma}^2 = (n - q - 2)^{-1} (Y - H\hat{\beta})^T A^{-1} (Y - H\hat{\beta})$. This posterior covariance term gives information about the difference in form between the mean posterior and the mean prior functions. The first term within the brackets on the right side of the equation, $\chi(x_1, x_2)$, represents the correlation function dependent upon the different inputs. The second term, $t(x_1)^T A^{-1} t(x_2)$, arises from the correlation of $\mathbf{Y}$ at an input location and its associated predicted emulator outcome with the training set. The third term, $(h(x_1)^T - t(x_1)^T A^{-1} H) * (H^T A^{-1} H)^{-1} (h(x_2)^T - t(x_2)^T A^{-1} H)^T$ is a covariance quantity related to the residuals from the mean posterior function, the regression function.

References

- Alves O, Wang G, Zhong A, Smith N, Tseitkin F, Warren G., Schiller A, Godfrey S, Meyers G (2003) POAMA: Bureau of Meteorology operational coupled model seasonal forecast system. In: Stone R, Partridge I (eds) Science for drought: proceedings of the national drought forum, Department of Primary Industries, Brisbane, April 2003, pp 49–56
- Balmaseda MA, Mogensen K, Weaver AT (2013) Evaluation of the ecmwf ocean reanalysis system oras4. *QJR Meteorol Soc* 139:1132–1161. <https://doi.org/10.1002/qj.2063>
- Barnston AG, Livezey RE (1987) Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Mon Weather Rev* 115(6):1083–1126. [https://doi.org/10.1175/1520-0493\(1987\)115<1083:CLSP>2.0.CO;2](https://doi.org/10.1175/1520-0493(1987)115<1083:CLSP>2.0.CO;2)
- Bitz CM, Lipscomb WH (1999) An energy-conserving thermodynamic model of sea ice. *J Geophys Res* 104:15,669–15,677
- Branstator G, Teng H, Meehl GA, Kimoto M, Knight JR, Latif M, Rosati A (2012) Systematic estimates of initial-value decadal predictability for six aogcms. *J Climate* 25(6):1827–1846
- Carton JA, Giese BS (2008) A reanalysis of ocean climate using simple ocean data assimilation (soda). *Mon Weather Rev* 136(8):2999–3017
- Challenor P, McNeill D, Gattiker J (2010) Assessing the probability of rare climate events. In: O'Hagan A, West M (eds) The Oxford handbook of applied Bayesian analysis. Oxford University Press, Oxford, pp 403–430
- Chang Y, Zhang S, Rosati A, Delworth T, Stern W (2013) An assessment of oceanic variability for 1960–2010 from the gfdl ensemble coupled data assimilation. *Clim Dyn* 40(3–4):775–803. <https://doi.org/10.1007/s00382-012-1412-2>
- Collins W, Bitz C, Blackmon M, Bonan G, Bretherton C, Carton J, Chang P, Doney S, Hack J, Henderson T, Kiehl J, Large W, McKenna D, Santer B, Smith R (2006) The community climate system model: Ccsm3. *J Clim* 19:2122–2143. <https://doi.org/10.1029/GL021592>
- Goddard L, Kumar A, Solomon A, Smith D, Boer G, Gonzalez P, Kharrin V, Merryfield W, Deser C, Mason SJ et al (2013) A verification framework for interannual-to-decadal predictions experiments. *Clim Dyn* 40(1–2):245–272. <https://doi.org/10.1007/s00382-012-1481-2>
- Holden PB, Edwards NR (2010) Dimensionally reduced emulation of an AOGCM for application to integrated assessment modelling. *Geophys Res Lett* 37:L21707. <https://doi.org/10.1029/2010GL045137>
- Hong SY, Koo MS, Jang J, Kim JE, Park H, Joh M, Kang JH, Oh TJ (2013) An evaluation of the software system dependency of a global atmospheric model. *Mon Weather Rev* 141:4165–4172
- Hotelling H (1933) Analysis of a complex of statistical variables with principal components. *J Educational Psychol* 24:417–441
- Hunke EC, Dukowicz JK (2003) The sea ice momentum equation in the free drift regime. Los Alamos National Laboratory, Tech. Rep. LA-UR-03-2219
- Kay J, Deser C, Phillips A, Mai A, Hannay C, Strand G, Arblaster J, Bates S, Danabasoglu G, Edwards J et al (2015) The community earth system model (cesm) large ensemble project: a community resource for studying climate change in the presence of internal climate variability. *Bull Am Meteorol Soc* 96(8):1333–1349
- Köhl A, Stammer D (2008) Variability of the meridional overturning in the north atlantic from the 50-year gecco state estimation. *J Phys Oceanogr* 38(9):1913–1930
- Kröger J, Müller WA, von Storch JS (2012) Impact of different ocean reanalyses on decadal climate prediction. *Clim Dyn* 39(3–4):795–810
- Large W, Yeager S (2009) The global climatology of an interannually varying air-sea flux data set. *Clim Dyn* 33(2–3):341–364
- Li TY, Yorke JA (1975) Period three implies chaos. *Am Math Mon* 82:985–992
- Lorenz EN (1963) Deterministic nonperiodic flow. *J Atmos Sci* 20(2):130–141. [https://doi.org/10.1175/1520-0469\(1963\)020<0130:DNF>2.0.CO;2](https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2)
- Matei D, Pohlmann H, Jungclaus J, Müller W, Haak H, Marotzke J (2012) Two tales of initializing decadal climate prediction experiments with the echam5/mpi-om model. *J Clim* 25(24):8502–8523
- McKay MD, Beckman RJ, Conover WJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21(2):239–245. <http://www.jstor.org/stable/1268522>
- Meehl GA, Covey C, Delworth T, Latif M, McAvaney B, Mitchell JFB, Stouffer RJ, Taylor KE (2007) The wcrp cmip3 multi-model dataset: a new era in climate change research. *Bull Am Meteorol Soc* 88:1383–1394
- Pearson K (1901) Liii. on lines and planes of closest fit to systems of points in space. *Lond Edinburgh Dublin Philosophical Mag J Sci* 2(11):559–572. <https://doi.org/10.1080/14786440109462720>
- Pohlmann H, Smith DM, Balmaseda MA, Keenlyside NS, Masina S, Matei D, Müller WA, Rogel P (2013) Predictability of the mid-latitude atlantic meridional overturning circulation in a multi-model system. *Clim Dyn* 41(3–4):775–785. <https://doi.org/10.1175/JCLI-D-11-00633.1>
- Rasmussen C, Williams C (2006) Gaussian processes for machine learning. MIT Press, Cambridge, MA
- Rougier J, Sexton DM (2007) Inference in ensemble experiments. *Philosophical Trans R Soc Lond A* 365(1857):2133–2143
- Schrier G, Maas L (1998) Chaos in a simple model of the three-dimensional, salt-dominated ocean circulation. *Clim Dyn* 14:489–502. <https://doi.org/10.1007/s003820050236>
- Smith DM, Cusack S, Colman AW, Folland CK, Harris GR, Murphy JM (2007) Improved surface temperature prediction for the coming decade from a global climate model. *Science* 317(5839):796–799. <https://doi.org/10.1126/science.1139540>

- Smith DM, Scaife AA, Kirtman BP (2012) What is the current state of scientific knowledge with regard to seasonal and decadal forecasting? *Environ Res Lett* 7(1):015,602. <https://doi.org/10.1088/1748-9326/7/1/015602>
- Smith RD, Gent PR (2002) Reference manual for the parallel ocean program (pop), ocean component of the community climate system model (ccsm2.0 and 3.0). Technical Report LA-UR-02-2484
- Striver RL, Forest CE, Keller K (2015) Effects of initial conditions uncertainty on regional climate variability: An analysis using a low-resolution cesm ensemble. *Geophys Res Lett* 42(13):5468–5476. <https://doi.org/10.1002/2015GL064546>
- Taylor K, Stouffer R, Meehl GA (2012) An overview of cmip5 and the 1205 experiment design. *J Atmosph/Ocean Tech* 93:1704–1715
- Tokmakian R, Challenor P (2014) Uncertainty in modeled upper ocean heat content change. *Clim Dyn* 42:1–20
- Tokmakian R, Challenor P, Andrianakis I (2012) An extreme non-linear example of the use of emulators with simulators using the stommel model. *J Atmos and Oceanic Tech* 29:1704–1715
- Trenberth K (1997) The definition of el niño. *Bull Am Meteorol Soc* 78(12):2771–2778. <https://doi.org/10.1175/1520-0477>
- Troccoli A, Palmer T (2007) Ensemble decadal predictions from analysed initial conditions. *Philosophical Trans R Soc Lond A* 365(1857):2179–2191. <https://doi.org/10.1098/rsta.2007.2079>
- Urban NM, Fricker TE (2010) A comparison of latin hypercube and grid ensemble designs for the multivariate emulation of an earth system model. *Comput Geosci* 36:746–755. <https://doi.org/10.1016/j.cageo.2009.11.004>
- Valdivieso M, Haines K, Zuo H (2012) Myocean scientific validation report (scvr) for v2.1 reprocessed analysis and reanalysis. Tech. Rep. WP04 GLO U Reading v2.1, GMES Marine Core Services
- Yeager S, Karspeck A, Danabasoglu G, Tribbia J, Teng H (2012) A decadal prediction case study: late twentieth-century north atlantic ocean heat content. *J Clim* 25(15):5173–5189

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.