Uncertainty Quantification of the Coupled Ocean-Atmosphere System
Gino Chen

Introduction

All global models have difficulties capturing the real world signal due to chaotic effects and model errors. However, the statistics (signal and noise) will be significantly improved by using multiple model ensemble (MME) forecast under suitable conditions. A single model ensemble run is not able to capture the statistics as does the MME forecast. This indicates that there exists a "model discrepancy" by using a single model, which can be attributed to failure in capturing the "sufficient dynamics" the multi-model possessed. We hypothesize that the model discrepancy could be understood from the statistics of the MME state output. The MME state of the process (initial or oscillatory states) will determine how the distribution of the parameter uncertainties behave. This raises the question, can we reduce the numbers of physical models to preform the same statistics as the MME seasonal forecast by treating multi-model run as the "true observation"?

Hence, we could start by arbitrarily perturbing the model parameters we have physically chosen. This will give us a clue for where the uncertainties dwell; which parameters are most sensitive to small perturbations. Eventually use the model error and measurement error to correct the shape of the parameter distribution. We will analyze outputs of the MME runs and see how the state probability distribution evolves throughout the process. Our goal is to treat the MME runs as the "true observations/realistic data", and try to introduce uncertainties to a single model to explore what the "realistic data" encompassed, eventually quantify the uncertainties to the best.

Data and Methodology

We will use a toy model to mimic the MME ENSO prediction statistics. This simple model is a slow and a fast timescale Lorenz model coupled systems of equations. It has the feature of a chaotic atmosphere and a slow evolving ocean component. We will generate a "realistic run" according to a fixed parameter, and introduce the error statistics of MME at the specified time sample. This model does not include spacial dimensions, so we consider only the NINO3.4 index time series as our forecast target. Hence, the only sample index is the time sample.

Coupled Lorenz Equation:

\[
\begin{align*}
\frac{dx}{dt} &= \sigma(y_a - x_a) - \alpha(x_o + k1) \\
\frac{dy}{dt} &= rx_a - y_a - x_ao + \alpha(y_o + k1) \\
\frac{dz}{dt} &= x_o y_a - bz_o + \alpha z_o \\
\frac{dx_o}{dt} &= \tau[\sigma(y_o - x_o) - \frac{\alpha}{\tau}(x_o + k1)] \\
\frac{dy_o}{dt} &= \tau[rx_o - y_o - xoao + \frac{\alpha}{\tau}(y_o + k1)] \\
\frac{dz_o}{dt} &= \tau[x_o yo - bzo + \frac{\alpha}{\tau}z_o]
\end{align*}
\]

Uncertain parameters of this system:

\(\alpha\): coupling strength
\(\tau\): ocean timescale
\(k_1\): offset parameter
\(\sigma, b, r\): standard values of Lorenz 63 model

The approach to quantify the model parameter uncertainty is to use Polynomial Chaos Expansion (PCE) to generate a "surrogate" of the output, and find the statistics based on this surrogate. The general approach, which is much more costly, is by re-running the model over and over again on different uncertain parameters to determine the model statistics. The PCE ideally explores the entire parameter space efficiently by projecting the solutions onto a few orthogonal basis, according to the probability weighting function, in parameter space to get the PCE coefficients and build truncated series expansion based on these results. This technique allows us to physically determine which parameters are most sensitive to the forecast, and will give us an idea where most of the uncertainty of the forecasts comes from. Which leads us to filtering out the insensitive parameters in the future which will enhance the computation efficiency.

PC expansion has the solution expanded as the infinite stochastic basis in the form:

\[U(t, \theta) = \sum_{k=0}^{\infty} \hat{u}_k(t) \psi_k(\theta)\]

The probability inner product of two vectors,

\[E[uv] = \langle u, v \rangle = \int_{\Theta} u(\theta) v(\theta) p(\theta) d\theta\]

This guarantees the orthogonality of the basis in \(L_2(\Theta, p(\theta))\) space. Hence we get the PC coefficient by projecting the solutions onto the stochastic basis:

\[\hat{u} = \frac{\langle U, \psi_k \rangle}{\langle \psi_k, \psi_k \rangle} = \frac{E[U \psi_k]}{E[\psi_k \psi_k]}, \quad k = 0, 1, ...\]

We use Non-intrusive Spectral Projection (NISP) to avoid changing the codes of the model, and treat the model as a black box while inputting uncertainties prior to running the model. The spectral projection will be treated after the model run which greatly reduced the complication of doing an Intrusive Method.

We design a Bayesian Hierarchical Model to further correct the distribution of this uncertainty input. At this stage we assume that we have already explored most of the possible range of uncertainties this model could capture with phenomenon we are interested in. The next step is to correct the "physical bias" a single model misses the state of the process. Such as the weaker and skewed variance spread at different natural oscillation phases compared to the MME runs. Also this leads to multiple lead time startup schemes for the seasonal forecast, since there is a bias for different lead time initial memory (figure out why there is a bias and how are these bias behaving).

The dependencies of the realistic data and model run could be written out as:

\[d_i = M_i(\theta) + E' + E''\]

\(i\): time sampling index
\(d\): observational data
\(M\): model output
\(\theta\): uncertain parameter input
\(E'\): measurement error (assumed to be fixed in time)
\(E''\): model error (independent of parameters)

The measurement errors are assumed to be normally distributed with a zero mean and constant variance chosen empirically, \(E' \sim N(0, \sigma^2)\). Whereas the model error is assumed...
to be normally distributed with zero mean and time-varying variance according to different phases of the process, \( E_i \sim N(0, \sigma_i^2) \).

Bayesian Inference: \( p(\theta|d) = \frac{p(d|\theta)p(\theta)}{p(d)} = \frac{p(d|\theta)p(\theta)}{\int p(d|\theta)p(\theta)d\theta} \)

Since the error densities are independent of \( \theta \), we get

\[ p(E|\theta^i) = p(E) \]

This shows that for a given \( \theta \), all the uncertainties of the data comes from the error distribution

\[ p(d|\theta^i) = p(E|\theta^i) = p(E) = p(E' + E''_i) = p(d - M(\theta^i)) \]

This gives a multivariate normal distribution for all the sampling data also called likelihood function

\[ p(d|\theta^i) = L(\theta^i) = \prod_{i=1}^{n} p(d_i - M_i(\theta^i)) \propto \prod_{i=1}^{n} p(E' + E''_i) \propto \]

\[ \prod_{i=1}^{n} e^{-\frac{(d_i - M_i(\theta^i))}{\sigma_i^2}} \]

Since the denominator of Bayesian Inference is just a scaling factor for fixed data, we know

\[ p(\theta|d) \propto L(\theta)p(\theta) \]

This is the posterior density we use to estimate the statistics of the single model ensemble forecast. For instance, the expectation for a variable \( f(t) \):

\[ E[f(t)] = \int_\theta f(t, \theta)L(\theta)p(\theta)d\theta = \int_\theta f(t, \theta)p(\theta|d)d\theta \]

Since this is integrated through all possible prior support, we could reduce the integration cost greatly by sampling on the posterior support, where the density peaks the most, by doing MCMC sampling. We will follow Metropolis-Hastings Random Walk Sampling Scheme for the time being. This will give the discrete integration:

\[ E[f(t)] = \frac{1}{n-b} \sum_{i=b+1}^{n} f_i(t) \]

b: MCMC burnin numbers
n: MCMC total iterations

Once the parameter distributions are corrected by Bayesian inference, and we have the sampling densities at hand, we could use the surrogate without re-running the model and see whether the statistics are capturing the MME characteristics.

We apply PCE, Bayesian Inference and MCMC to the coupled Lorenz model and design a ”true run” that modifies a parameter which the PC expansion does not include as a model error. At the same time include the characteristic spread error in the ”true run” as an additional model error that a single model ensemble is missing to the MME. Then try to use the residual parameters to capture the signal and spread of the ”truth”. From Figure 1, we could see the spread for the Monte-Carlo model run not so strong compared to the ”observation spread”. After the probability inversion the expected results show a stronger spread throughout different phases. The parameter range shown in Figure 2 shows a wider spread of ‘sigma’ and ‘k’, which might indicate these parameters are trying to capture the spread effect. The narrow spread of ‘tau’ might indicate the higher sensitivity to small perturbation which makes it tend to stay at a narrow range.

Future Work

The spread error is not well captured in this test, so the priority is on the designing of the hierarchical model to capture the spread error. Test the convergence of sampling by Kullback-Leibler divergence (KLD), quantify the difference between the exact posterior and approximated posterior.

Reference
