Data Assimilation and Its Connections with Uncertainty Quantification, Forecast and Machine Learning

Nan Chen

Department of Mathematics
University of Wisconsin-Madison

New York Scientific Data Summit 2021
October 26-29, 2021

Outline

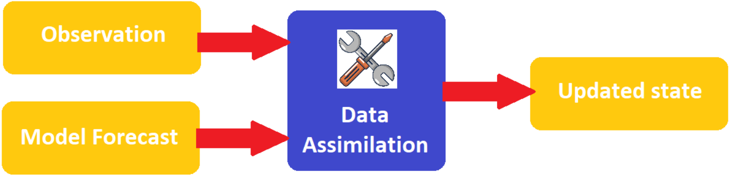
- Introduction to Data Assimilation

- Data Assimilation and Uncertainty Quantification

- Data Assimilation and Solving High-Dimensional Fokker-Planck Equations

- Data Assimilation and Machine Learning Forecast

Introduction to Data Assimilation



Observation

Model Forecast

Data Assimilation

Updated state

Data Assimilation (DA)

Data assimilation (DA) seeks to optimally combine a numerical model with observations to improve results, where

- the model is typically chaotic and has uncertainties, while
- the observations contain noise and are often available only for a subset of the state variables.

(Kalnay 2003; Majda & Harlim 2012; Law, Stuart & Zygalakis, 2015 ...)

Data Assimilation (DA)

Data assimilation (DA) seeks to optimally combine a numerical model with observations to improve results, where

- the model is typically chaotic and has uncertainties, while
- the observations contain noise and are often available only for a subset of the state variables.

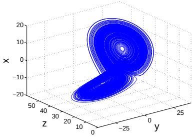
(Kalnay 2003; Majda & Harlim 2012; Law, Stuart & Zygalakis, 2015 ...)

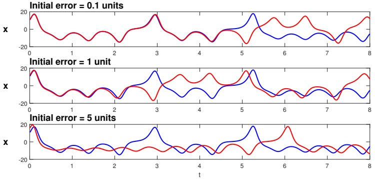
DA has wide applications in many areas, such as engineering, climate science, geophysics, neural science and material science.

When the underlying system is chaotic or turbulent, DA plays an extremely important role in state estimation and improving the initialization for forecast.

Example:
The Lorenz 63 model

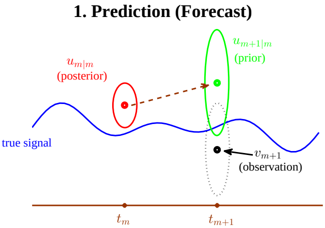
$$
\begin{aligned}
dx &= \sigma(y-x)dt, \\
dy &= \left(x(\rho-z)-y\right)dt, \\
dz &= (xy-\beta z)dt.
\end{aligned}
$$

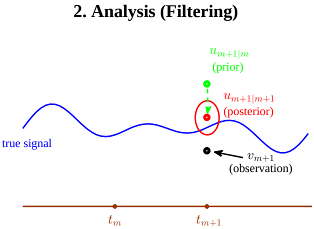




- Underlying principle of DA: **the Bayes' theorem**

$$\underbrace{p(u_{m+1}|v_{m+1})}_{\text{posterior}} \sim \underbrace{p(u_{m+1})}_{\text{prior}}\underbrace{p(v_{m+1}|u_{m+1})}_{\text{likelihood}}.$$

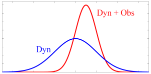


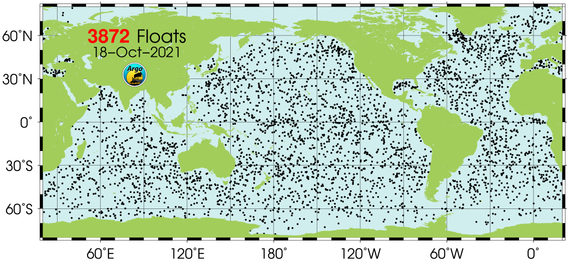


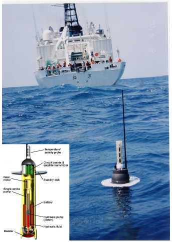
Data Assimilation and Uncertainty Quantification

Recovering Geophysical Flows with Lagrangian Data Assimilation

- Lagrangian tracers: drifters/floaters following a parcel of fluid's movement.
- **[Inverse Problems].** Data assimilation with Lagrangian tracers: recovering the underlying velocity field with observations (from tracers).
  - Only dynamics: large uncertainty due to turbulence.
  - Dynamics + Observations: reducing error and uncertainty.







(From: UCSD ARGO program)

Lagrangian data assimilation is a hot topic recently with a wide range of applications.

An important question:

What is the uncertainty reduction as a function of the number of tracers?

Model set-up for studying the uncertainty reduction v.s. the number of tracers.

1. Underlying flow model.

Consider a random flow modeled by a finite number of Fourier modes with random amplitudes in double periodic domain $(0, 2\pi]^2$,

$$\vec{v}(\vec{x}, t) = \sum_{\vec{k} \in \mathbf{K}} \hat{v}_{\vec{k}}(t) \cdot e^{i\vec{k}\cdot\vec{x}} \cdot \vec{r}_{\vec{k}}.$$

Each $\hat{v}_{\vec{k}}(t)$ follows an Ornstein-Uhlenbeck (O.U.) process,

$$d\hat{v}_{\vec{k}}(t) = -d_{\vec{k}}\hat{v}_{\vec{k}}(t)dt + f_{\vec{k}}(t)dt + \sigma_{\vec{k}}dW_{\vec{k}}^{v}(t).$$

Model set-up for studying the uncertainty reduction v.s. the number of tracers.

1. Underlying flow model.

Consider a random flow modeled by a finite number of Fourier modes with random amplitudes in double periodic domain $(0, 2\pi]^2$,

$$\vec{v}(\vec{x}, t) = \sum_{\vec{k}\in\mathbf{K}} \hat{v}_{\vec{k}}(t) \cdot e^{i\vec{k}\cdot\vec{x}} \cdot \vec{r}_{\vec{k}}.$$

Each $\hat{v}_{\vec{k}}(t)$ follows an Ornstein-Uhlenbeck (O.U.) process,

$$d\hat{v}_{\vec{k}}(t) = -d_{\vec{k}}\hat{v}_{\vec{k}}(t)dt + f_{\vec{k}}(t)dt + \sigma_{\vec{k}}dW^{v}_{\vec{k}}(t).$$

2. Observations.

The observations are given by the trajectories of L noisy Lagrangian tracers,

$$\begin{aligned} d\vec{x}_l(t) &= \vec{v}(\vec{x}_l(t), t)dt + \sigma_x dW^x_l(t) \\ &= \sum_{\vec{k}\in\mathbf{K}} \underbrace{\hat{v}_{\vec{k}}(t) \cdot e^{i\vec{k}\cdot\vec{x}_l(t)} \cdot \vec{r}_{\vec{k}}}_{\text{Nonlinear!}} dt + \sigma_x dW^x_l(t), \quad l = 1, \ldots, L. \end{aligned}$$

Model set-up for studying the uncertainty reduction v.s. the number of tracers.

1. Underlying flow model.

Consider a random flow modeled by a finite number of Fourier modes with random amplitudes in double periodic domain $(0, 2\pi]^2$,

$$\vec{v}(\vec{x}, t) = \sum_{\vec{k}\in\mathbf{K}} \hat{v}_{\vec{k}}(t) \cdot e^{i\vec{k}\cdot\vec{x}} \cdot \vec{r}_{\vec{k}}.$$

Each $\hat{v}_{\vec{k}}(t)$ follows an Ornstein-Uhlenbeck (O.U.) process,

$$d\hat{v}_{\vec{k}}(t) = -d_{\vec{k}}\hat{v}_{\vec{k}}(t)dt + f_{\vec{k}}(t)dt + \sigma_{\vec{k}}dW^v_{\vec{k}}(t).$$

2. Observations.

The observations are given by the trajectories of L noisy Lagrangian tracers,

$$\begin{aligned} d\vec{x}_l(t) &= \vec{v}(\vec{x}_l(t), t)dt + \sigma_x dW_l^x(t) \\ &= \sum_{\vec{k}\in\mathbf{K}} \underbrace{\hat{v}_{\vec{k}}(t) \cdot e^{i\vec{k}\cdot\vec{x}_l(t)} \cdot \vec{r}_{\vec{k}}}_{\text{Nonlinear!}} dt + \sigma_x dW_l^x(t), \qquad l = 1, \ldots, L. \end{aligned}$$

3. Combining model and observations: A nonlinear DA framework — $p(\mathbf{U}|\mathbf{X})$.

Observations:	$d\mathbf{X} = \mathbf{P}_X(\mathbf{X})\mathbf{U}dt + \Sigma_x dW_X,$	$\mathbf{X} = (x_{1,x}, x_{1,y}, ..., x_{L,x}, x_{L,y})^T,$
Underlying flow:	$d\mathbf{U} = -\Gamma\mathbf{U}dt + \mathbf{F}(t)dt + \Sigma_u dW_u,$	$\mathbf{U} = (\hat{v}_1, ...\hat{v}_{\mathbf{K}})^T.$

Closed analytic formulae are available for such a nonlinear DA (Chen & Majda 2018).

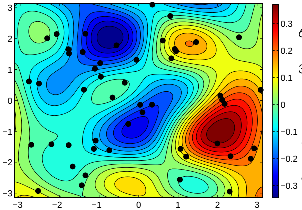
1. Recovering random incompressible flows
First rigorous math theory

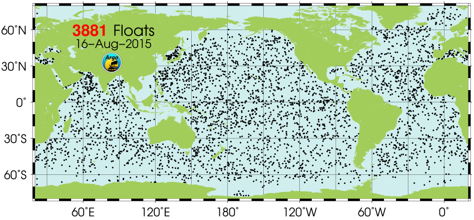
(Chen, Majda & Tong, *Nonlinearity*, 2014)

Prior distribution	based only on the model	$p(\mathbf{U}_t) \sim \mathcal{N}(\mathbf{m}_t^{att}, \mathbf{R}_t^{att})$
Posterior distribution	combining model and obs	$p(\mathbf{U}_t \vert \mathbf{X}_{s\leq t}) \sim \mathcal{N}(\mathbf{m}_t, \mathbf{R}_t)$

Theorem (Invariant measure of tracers)

The distribution of the noisy tracers $\vec{x}_l(s)$ converges geometrically fast towards the uniform distribution on $(0, 2\pi]^2$.





Theorem (Asymptotic behavior of the posterior statistics)

The posterior mean converges to the truth while the posterior covariance decreases as a function of $L^{-1/2}$.

To quantify the uncertainty reduction in the posterior distribution $p(\mathbf{U}_t|\mathbf{X}_{s\le t})$ related to the prior $p(\mathbf{U}_t)$, an **information criterion** — the *relative entropy* — is adopted:

$$\mathcal{P}(p(\mathbf{U}_t|\mathbf{X}_{s\le t}), p(\mathbf{U}_t)) = \int p(\mathbf{U}_t|\mathbf{X}_{s\le t}) \ln \frac{p(\mathbf{U}_t|\mathbf{X}_{s\le t})}{p(\mathbf{U}_t)}$$

For Gaussian distributions,

$$\begin{aligned}
&\mathcal{P}(p(\mathbf{U}_t|\mathbf{X}_{s\le t}), p(\mathbf{U}_t)) \\
&\quad = \frac{1}{2}\left[(\mathbf{m}_t - \mathbf{m}_t^{att})^*(\mathbf{R}_t^{att})^{-1}(\mathbf{m}_t - \mathbf{m}_t^{att})\right] && \cdots \text{Signal} \\
&\quad + \frac{1}{2}\left[\mathrm{tr}(\mathbf{R}_t(\mathbf{R}_t^{att})^{-1}) - |\mathbf{K}| - \ln\det(\mathbf{R}_t(\mathbf{R}_t^{att})^{-1})\right] && \cdots \text{Dispersion}
\end{aligned}$$

To quantify the uncertainty reduction in the posterior distribution $p(\mathbf{U}_t|\mathbf{X}_{s\leq t})$ related to the prior $p(\mathbf{U}_t)$, an **information criterion** — the *relative entropy* — is adopted:

$$\mathcal{P}(p(\mathbf{U}_t|\mathbf{X}_{s\leq t}), p(\mathbf{U}_t)) = \int p(\mathbf{U}_t|\mathbf{X}_{s\leq t}) \ln \frac{p(\mathbf{U}_t|\mathbf{X}_{s\leq t})}{p(\mathbf{U}_t)}$$

For Gaussian distributions,

$$\begin{aligned}
&\mathcal{P}(p(\mathbf{U}_t|\mathbf{X}_{s\leq t}), p(\mathbf{U}_t)) \\
&=\frac{1}{2}\left[(\mathbf{m}_t - \mathbf{m}_t^{att})^*(\mathbf{R}_t^{att})^{-1}(\mathbf{m}_t - \mathbf{m}_t^{att})\right] & \cdots\text{Signal} \\
&+\frac{1}{2}\left[\operatorname{tr}(\mathbf{R}_t(\mathbf{R}_t^{att})^{-1}) - |\mathbf{K}| - \ln\det(\mathbf{R}_t(\mathbf{R}_t^{att})^{-1})\right] & \cdots\text{Dispersion}
\end{aligned}$$

Theorem (Uncertainty Reduction)

As $L \to \infty$, there exists a fixed time $s_0 > 0$ such that for a.s. $\vec{v}_{s\leq t}$

$$\text{For any } t > s_0, \qquad \text{Signal} \to \frac{1}{2}(\mathbf{U}_t - \mathbf{m}_t^{att})^*\mathbf{R}_{att}^{-1}(\mathbf{U}_t - \mathbf{m}_t^{att}) \ \text{ in } \ \mathbf{P}_{\vec{v}_{s\leq t}},$$

$$\text{For any } t > 0, \qquad \frac{\text{Dispersion}}{\frac{|\mathbf{K}|+2}{4}\ln L} \to 1 \ \text{ in } \ \mathbf{P}_{\vec{v}_{s\leq t}}.$$

Reducing the uncertainty by a fixed amount requires an exponential increase in the number of tracers — **A practical information barrier**!

To quantify the uncertainty reduction in the posterior distribution $p(\mathbf{U}_t|\mathbf{X}_{s\le t})$ related to the prior $p(\mathbf{U}_t)$, an **information criterion** — the *relative entropy* — is adopted:

$$\mathcal{P}(p(\mathbf{U}_t|\mathbf{X}_{s\le t}), p(\mathbf{U}_t)) = \int p(\mathbf{U}_t|\mathbf{X}_{s\le t}) \ln \frac{p(\mathbf{U}_t|\mathbf{X}_{s\le t})}{p(\mathbf{U}_t)}$$

For Gaussian distributions,

$$\begin{aligned}&\mathcal{P}(p(\mathbf{U}_t|\mathbf{X}_{s\le t}), p(\mathbf{U}_t))\\ &=\frac{1}{2}\left[(\mathbf{m}_t-\mathbf{m}_t^{att})^*(\mathbf{R}_t^{att})^{-1}(\mathbf{m}_t-\mathbf{m}_t^{att})\right] && \cdots\text{Signal}\\ &+\frac{1}{2}\left[\mathrm{tr}(\mathbf{R}_t(\mathbf{R}_t^{att})^{-1})-|\mathbf{K}|-\ln\det(\mathbf{R}_t(\mathbf{R}_t^{att})^{-1})\right] && \cdots\text{Dispersion}\end{aligned}$$

Theorem (Uncertainty Reduction)

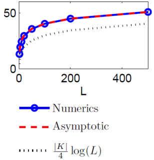
As $L\to\infty$, there exists a fixed time $s_0>0$ such that for a.s. $\vec{v}_{s\le t}$

For any $t>s_0$, $\quad$ Signal $\to \frac{1}{2}(\mathbf{U}_t-\mathbf{m}_t^{att})^*\mathbf{R}_{att}^{-1}(\mathbf{U}_t-\mathbf{m}_t^{att})$ in $\mathbf{P}_{\vec{v}_{s\le t}}$,

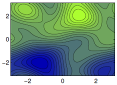
For any $t>0$, $\quad \dfrac{\text{Dispersion}}{\frac{|\mathbf{K}|+2}{4}\ln L}\to 1$ in $\mathbf{P}_{\vec{v}_{s\le t}}$.

Reducing the uncertainty by a fixed amount requires an exponential increase in the number of tracers — **A practical information barrier**!

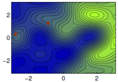
Uncertainty reduction in dispersion



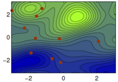
Truth



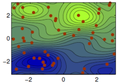
L=2



L=10



L=50



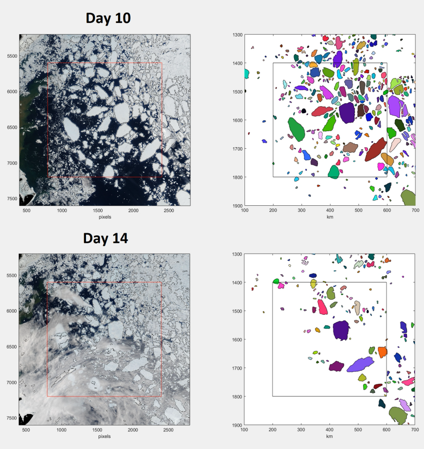
2. Noisy Lagrangian tracers for recovering random rotating compressible flows

(Chen, Majda & Tong, *JNLS* 2015; Chen & Majda, *MWR*, 2016)

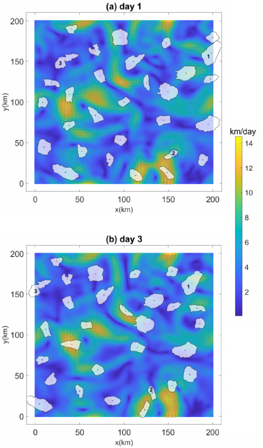
- Underlying flow field is multiscale, containing the slow geostrophically balanced (GB) modes and fast gravity modes.
- Highly nonlinear observations mixing (GB) and gravity modes!
- Designed several cheap reduced DA strategies, which have comparable high skill in recovering GB modes as using the full system, in the geophysical scenario with small Rossby number.

3. Lagrangian data assimilation using the observed sea ice floes.

Satellite images: the marginal ice zone, June 2008.



Model simulations.



- Developed a coupled ocean, atmosphere and discrete element sea ice model.
- Developed a cheap DA to recover the ocean current beneath the sea ice floes.
- Developed a DA-based dynamical interpolation to recover the missing obs of the sea ice floes in the presence of clouds, applying to the satellite images. (Chen, Fu & Manucharyan, 2021; Covington, Chen, Wilhelmus, & Lopez, 2021)

Data Assimilation and Solving High-Dimensional Fokker-Planck Equations

Consider a general nonlinear dynamical system with noise,

$$d\mathbf{u} = \mathbf{F}(\mathbf{u}, t)dt + \mathbf{\Sigma}(\mathbf{u}, t)d\mathbf{W}.$$

The Fokker-Planck equation describes the time evolution of the probability density function (PDF) associated with **u**,

$$\frac{\partial}{\partial t} p(\mathbf{u}, t) = -\nabla_{\mathbf{u}}\big(\mathbf{F}(\mathbf{u}, t)p(\mathbf{u}, t)\big) + \frac{1}{2}\nabla_{\mathbf{u}} \cdot \nabla_{\mathbf{u}}(\mathbf{\Sigma}\mathbf{\Sigma}^T(\mathbf{u}, t)p(\mathbf{u}, t)).$$

Important applications:

- ensemble forecast
- linear response theory
- studying extreme events

Direct PDE solvers won't work efficiently for $dim > 3$. Monte Carlo simulations can handle slightly larger dimensional systems but still suffer from the **Curse of Dimensionality.**

Split the original systems into the following equivalent form:

$$
\begin{aligned}
d\mathbf{u}_{\mathbf{I}} &= \mathbf{F}_{\mathbf{I}}(t, \mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})dt + \boldsymbol{\Sigma}_{\mathbf{I}}(t, \mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})d\mathbf{W}_{\mathbf{I}}(t), \\
d\mathbf{u}_{\mathbf{II}} &= \mathbf{F}_{\mathbf{II}}(t, \mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})dt + \boldsymbol{\Sigma}_{\mathbf{II}}(t, \mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})d\mathbf{W}_{\mathbf{II}}(t),
\end{aligned}
$$

Split the original systems into the following equivalent form:

$$
\begin{aligned}
d\mathbf{u}_{\mathbf{I}} &= \mathbf{F}_{\mathbf{I}}(t, \mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})dt + \mathbf{\Sigma}_{\mathbf{I}}(t, \mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})d\mathbf{W}_{\mathbf{I}}(t),\\
d\mathbf{u}_{\mathbf{II}} &= \mathbf{F}_{\mathbf{II}}(t, \mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})dt + \mathbf{\Sigma}_{\mathbf{II}}(t, \mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})d\mathbf{W}_{\mathbf{II}}(t),
\end{aligned}
$$

Motivation: If we have an efficient DA scheme and L trajectories of $\mathbf{u}_{\mathbf{I}}$, namely $\mathbf{u}_{\mathbf{I}}^i$ with $i = 1, \ldots, L$, we can represent $p(\mathbf{u}_{\mathbf{II}}(t))$ by a summation of L conditional distributions,

$$
p(\mathbf{u}_{\mathbf{II}}(t)) = \lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} p\Big(\mathbf{u}_{\mathbf{II}}(t) | \mathbf{u}_{\mathbf{I}}^i(s \leq t)\Big).
$$

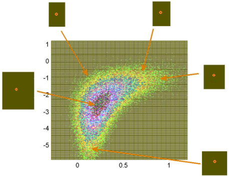
Split the original systems into the following equivalent form:

$$\begin{aligned} d\mathbf{u}_{\mathbf{I}} &= \mathbf{F}_{\mathbf{I}}(t, \mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})dt + \mathbf{\Sigma}_{\mathbf{I}}(t, \mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})d\mathbf{W}_{\mathbf{I}}(t), \\ d\mathbf{u}_{\mathbf{II}} &= \mathbf{F}_{\mathbf{II}}(t, \mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})dt + \mathbf{\Sigma}_{\mathbf{II}}(t, \mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})d\mathbf{W}_{\mathbf{II}}(t), \end{aligned}$$

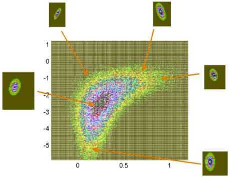
Motivation: If we have an efficient DA scheme and L trajectories of $\mathbf{u}_{\mathbf{I}}$, namely $\mathbf{u}_{\mathbf{I}}^i$ with $i = 1, \ldots, L$, we can represent $p(\mathbf{u}_{\mathbf{II}}(t))$ by a summation of L conditional distributions,

$$p(\mathbf{u}_{\mathbf{II}}(t)) = \lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} p\Big(\mathbf{u}_{\mathbf{II}}(t) | \mathbf{u}_{\mathbf{I}}^i(s \leq t)\Big).$$

Advantage: The method overcomes the curse of dimensionality! In other words, L does not increase as the dimension of $\mathbf{u}_{\mathbf{II}}$ increases.



MC



New method

The following nonlinear modeling framework includes a rich class of models

$$
\begin{aligned}
d\mathbf{u_I} &= [\mathbf{A}_0(t, \mathbf{u_I}) + \mathbf{A}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_I}(t, \mathbf{u_I})d\mathbf{W_I}(t), \\
d\mathbf{u_{II}} &= [\mathbf{a}_0(t, \mathbf{u_I}) + \mathbf{a}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_{II}}(t, \mathbf{u_I})d\mathbf{W_{II}}(t).
\end{aligned}
$$

Examples:

- physics-constrained nonlinear stochastic models

  (e.g., the noisy versions of Lorenz models, low-order models of Charney-DeVore flows, and a paradigm model for topographic mean flow interaction)
- stochastically coupled reaction-diffusion models in neuroscience and ecology

  (e.g., stochastically coupled FitzHugh-Nagumo models and stochastically coupled SIR epidemic models)
- multi-scale models in turbulence, fluids and geophysical flows

  (e.g., the Boussinesq equations with noise and stochastically forced rotating shallow water equation)

Many other models can be easily approximated by the above system.

The following nonlinear modeling framework includes a rich class of models

$$d\mathbf{u_I} = [\mathbf{A}_0(t, \mathbf{u_I}) + \mathbf{A}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_I}(t, \mathbf{u_I})d\mathbf{W_I}(t),$$
$$d\mathbf{u_{II}} = [\mathbf{a}_0(t, \mathbf{u_I}) + \mathbf{a}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_{II}}(t, \mathbf{u_I})d\mathbf{W_{II}}(t).$$

Examples:

- physics-constrained nonlinear stochastic models

  (e.g., the noisy versions of Lorenz models, low-order models of Charney-DeVore flows, and a paradigm model for topographic mean flow interaction)
- stochastically coupled reaction-diffusion models in neuroscience and ecology

  (e.g., stochastically coupled FitzHugh-Nagumo models and stochastically coupled SIR epidemic models)
- multi-scale models in turbulence, fluids and geophysical flows

  (e.g., the Boussinesq equations with noise and stochastically forced rotating shallow water equation)

Many other models can be easily approximated by the above system.

Key feature: Closed analytic formula is available for the nonlinear DA estimator $p\big(\mathbf{u_{II}}(t)|\mathbf{u}_\mathbf{I}^i(s \leq t)\big)$, which is a conditional Gaussian distribution.

The following nonlinear modeling framework includes a rich class of models

$$d\mathbf{u}_{\mathbf{I}} = [\mathbf{A}_0(t,\mathbf{u}_{\mathbf{I}}) + \mathbf{A}_1(t,\mathbf{u}_{\mathbf{I}})\mathbf{u}_{\mathbf{II}}]dt + \mathbf{\Sigma}_{\mathbf{I}}(t,\mathbf{u}_{\mathbf{I}})d\mathbf{W}_{\mathbf{I}}(t),$$
$$d\mathbf{u}_{\mathbf{II}} = [\mathbf{a}_0(t,\mathbf{u}_{\mathbf{I}}) + \mathbf{a}_1(t,\mathbf{u}_{\mathbf{I}})\mathbf{u}_{\mathbf{II}}]dt + \mathbf{\Sigma}_{\mathbf{II}}(t,\mathbf{u}_{\mathbf{I}})d\mathbf{W}_{\mathbf{II}}(t).$$

Examples:

- physics-constrained nonlinear stochastic models

  (e.g., the noisy versions of Lorenz models, low-order models of Charney-DeVore flows, and a paradigm model for topographic mean flow interaction)

- stochastically coupled reaction-diffusion models in neuroscience and ecology

  (e.g., stochastically coupled FitzHugh-Nagumo models and stochastically coupled SIR epidemic models)

- multi-scale models in turbulence, fluids and geophysical flows

  (e.g., the Boussinesq equations with noise and stochastically forced rotating shallow water equation)

Many other models can be easily approximated by the above system.

Key feature: Closed analytic formula is available for the nonlinear DA estimator $p\big(\mathbf{u}_{\mathbf{II}}(t)|\mathbf{u}_{\mathbf{I}}^i(s\leq t)\big)$, which is a conditional Gaussian distribution.

Assume the dimension of $\mathbf{u}_{\mathbf{I}}$ is low such that $p(\mathbf{u}_{\mathbf{I}}(t))$ can be approximated by a kernel density estimation. Then the time-dependent joint PDF is given by a Gaussian mixture,

$$p(\mathbf{u}_{\mathbf{I}}(t),\mathbf{u}_{\mathbf{II}}(t)) = \lim_{L\to\infty}\frac{1}{L}\sum_{i=1}^{L}\Big(K_{\mathbf{H}}(\mathbf{u}_{\mathbf{I}}(t)-\mathbf{u}_{\mathbf{I}}^i(t))\cdot p(\mathbf{u}_{\mathbf{II}}(t)|\mathbf{u}_{\mathbf{I}}^i(s\leq t))\Big).$$

Theorem: Error estimation (Chen, Majda & Tong *SIAM UQ* 2018).

Consider the following two ways of estimating the density p_t with the same L:

$\tilde{p}_t$: Kernel density estimation for the **joint PDF**.

$\hat{p}_t$: Hybrid method — Kernel density **+** conditional Gaussian estimation.

Results of the error estimates in light of the bias-variance decomposition:

$$\tilde{p}_t \text{ Bias bound} \geq \hat{p}_t \text{ Bias bound},$$

$$\frac{\tilde{p}_t \text{ Variance bound}}{\hat{p}_t \text{ Variance bound}} = \frac{H^{-\frac{N_{\mathbf{II}}}{2}} C}{\mathbb{E}\sqrt{\det(\mathbf{R}_{\mathbf{II}}(t))}^{-1}}.$$

Here $\mathbb{E}\sqrt{\det(\mathbf{R}_{\mathbf{II}}(t))}$ **does not decrease as L but the bandwidth H does!!**

When H shrinks and $N_{\mathbf{II}}$ becomes large, $H^{-\frac{N_{\mathbf{II}}}{2}}$ increases dramatically.

Equivalently, we have the following MISE estimations:

$$\tilde{p}_t: \ \text{MISE} \sim O\left(L^{-\frac{4}{4+N_{\mathbf{I}}+N_{\mathbf{II}}}}\right) \qquad \text{and} \qquad \hat{p}_t: \ \text{MISE} \sim O\left(L^{-\frac{4}{4+N_{\mathbf{I}}}}\right)$$

The error in the hybrid method **does not depend on $N_{\mathbf{II}}$**.

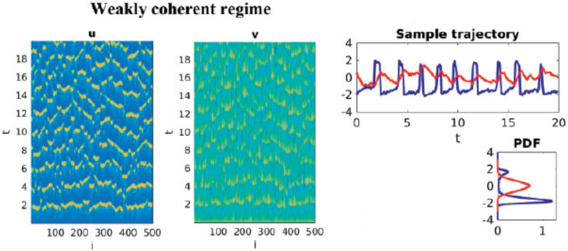
— Beating the curse of dimensionality in $\mathbf{u}_{\mathbf{II}}$!

Example: A nonlinear system with 1000 dimension: A Stochastic Coupled FHN Model (Chen & Majda 2017 PNAS).

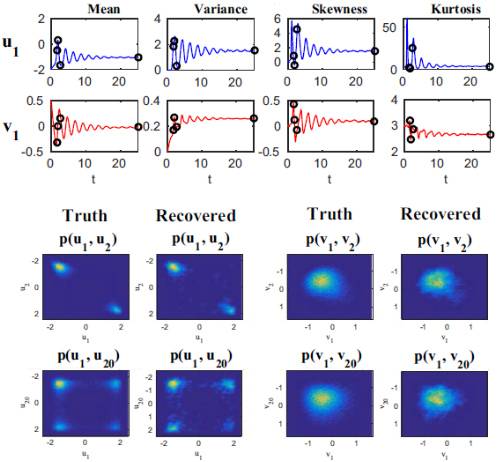
$$\epsilon d u_i = \left(d_u (u_{i+1} + u_{i-1} - 2 u_i) + u_i - \frac{1}{3} u_i^3 - v_i \right) dt + \sqrt{\epsilon} \delta_1 dW_{u_i},$$

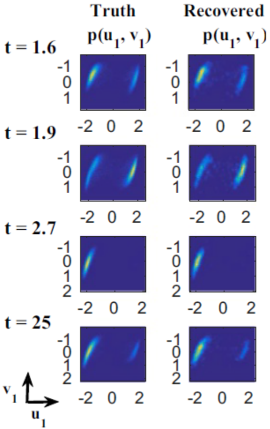
$$d v_i = (u_i + a) dt + \delta_2 dW_{v_i}, \qquad i = 1, \ldots, N.$$

with $N = 500$. The total number of dimension is 1000.



- Block decomposition allows an extremely efficient parallel computation of the covariance evolution.
- Statistical symmetry is incorporated and greatly reduces the number of sample L.
- An accurate recovery of both the transient and equilibrium non-Gaussian PDFs (one-point and two-point statistics) requires **only $L = 1$ samples**!
- As comparison, the truth is generated using Monte Carlo with $L_{MC} = 150,000$.





An accurate recovery of one-point and two-point statistics using only $L = 1$ samples!

Data Assimilation and Machine Learning Forecast

Model-Based Ensemble Forecast v.s. Machine Learning (ML) Forecast

Ensemble forecast based on physics-informed parametric models is one of the most widely used forecast algorithms for complex turbulent systems.

- **Major difficulty:** **model error**, ubiquitous in practice.

Model-Based Ensemble Forecast v.s. Machine Learning (ML) Forecast

Ensemble forecast based on physics-informed parametric models is one of the most widely used forecast algorithms for complex turbulent systems.

- **Major difficulty: model error**, ubiquitous in practice.

Nowadays, machine learning (ML) has become a powerful forecast tool.

- **Pros:** By extracting the information directly from the available observational data, the ML models can avoid the model error in the physics-informed models.
- **Cons:** only **partial, noisy, and possibly short observations** are available in many applications.

Model-Based Ensemble Forecast v.s. Machine Learning (ML) Forecast

Ensemble forecast based on physics-informed parametric models is one of the most widely used forecast algorithms for complex turbulent systems.

- **Major difficulty:** **model error**, ubiquitous in practice.

Nowadays, machine learning (ML) has become a powerful forecast tool.

- **Pros:** By extracting the information directly from the available observational data, the ML models can avoid the model error in the physics-informed models.
- **Cons:** only **partial, noisy, and possibly short observations** are available in many applications.

Can physics-informed models be combined with ML to improve the forecast?

Model-Based Ensemble Forecast v.s. Machine Learning (ML) Forecast

Ensemble forecast based on physics-informed parametric models is one of the most widely used forecast algorithms for complex turbulent systems.

- **Major difficulty:** **model error**, ubiquitous in practice.

Nowadays, machine learning (ML) has become a powerful forecast tool.

- **Pros:** By extracting the information directly from the available observational data, the ML models can avoid the model error in the physics-informed models.
- **Cons:** only **partial, noisy, and possibly short observations** are available in many applications.

Can physics-informed models be combined with ML to improve the forecast?

Yes!

Model-Based Ensemble Forecast v.s. Machine Learning (ML) Forecast

Ensemble forecast based on physics-informed parametric models is one of the most widely used forecast algorithms for complex turbulent systems.

- **Major difficulty: model error**, ubiquitous in practice.

Nowadays, machine learning (ML) has become a powerful forecast tool.

- **Pros:** By extracting the information directly from the available observational data, the ML models can avoid the model error in the physics-informed models.
- **Cons:** only **partial, noisy, and possibly short observations** are available in many applications.

Can physics-informed models be combined with ML to improve the forecast?

Yes! How? Via DA! DA can:

Model-Based Ensemble Forecast v.s. Machine Learning (ML) Forecast

Ensemble forecast based on physics-informed parametric models is one of the most widely used forecast algorithms for complex turbulent systems.

- **Major difficulty:** **model error**, ubiquitous in practice.

Nowadays, machine learning (ML) has become a powerful forecast tool.

- **Pros:** By extracting the information directly from the available observational data, the ML models can avoid the model error in the physics-informed models.
- **Cons:** only **partial, noisy, and possibly short observations** are available in many applications.

Can physics-informed models be combined with ML to improve the forecast?

Yes! How? Via DA! DA can:

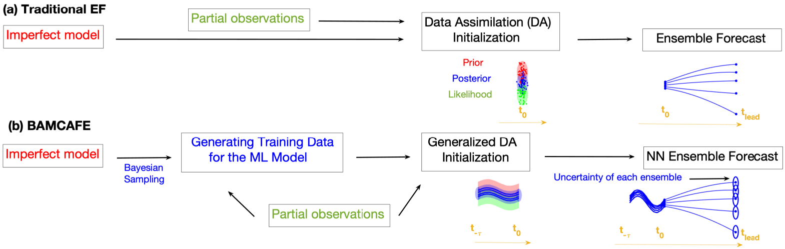
- mitigate the error in the time series generated from the parametric model,
- create the time series of the unobserved state variables, and
- generate multiple time series (via a Bayesian sampling) to provide enough training data for the ML forecast.

Assume we are only given **an imperfect/approximate model** and **partial and noisy observational time series**.

Bayesian Machine Learning Advanced Forecast Ensemble (BAMCAFE):

1. Generating the ML training data using a Bayesian sampling approach (i.e., a Bayesian ensemble DA).
2. Training a ML model (e.g., a neural network) utilizing the data from Step 1.
3. Employing a generalized DA for the initialization of the ML model.
4. Applying a ML ensemble forecast.

(See Chen & Li, *Chaos* 2021)



(a) Traditional EF

Imperfect model

Partial observations

Data Assimilation (DA) Initialization

Ensemble Forecast

(b) BAMCAFE

Imperfect model

Bayesian Sampling

Generating Training Data for the ML Model

Partial observations

Generalized DA Initialization

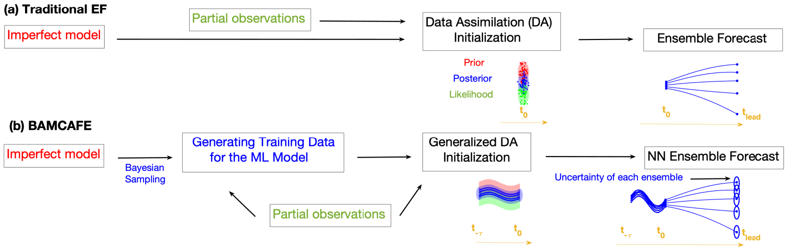
NN Ensemble Forecast

Assume we are only given **an imperfect/approximate model** and **partial and noisy observational time series**.

Bayesian Machine Learning Advanced Forecast Ensemble (BAMCAFE):

1. Generating the ML training data using a Bayesian sampling approach (i.e., a Bayesian ensemble DA).
2. Training a ML model (e.g., a neural network) utilizing the data from Step 1.
3. Employing a generalized DA for the initialization of the ML model.
4. Applying a ML ensemble forecast.

(See Chen & Li, *Chaos* 2021)



(a) Traditional EF

Imperfect model

Partial observations

Data Assimilation (DA) Initialization

Ensemble Forecast

(b) BAMCAFE

Imperfect model

Bayesian Sampling

Generating Training Data for the ML Model

Partial observations

Generalized DA Initialization

NN Ensemble Forecast

Uncertainty of each ensemble

In addition to forecasting the optimal point-wise value, the BAMCAFE also aims at providing an accurate quantification of the forecast uncertainty utilizing a non-Gaussian PDF constructed by a mixture distribution.

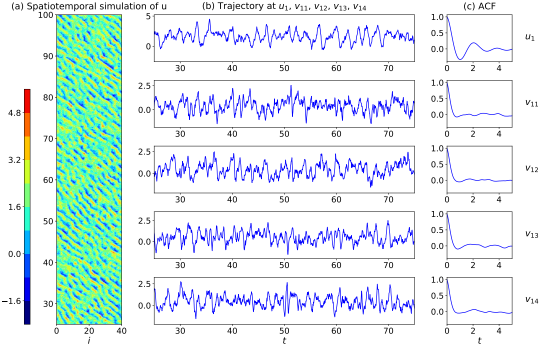
Numerical example.

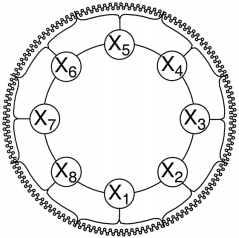
Perfect model: The two-layer Lorenz 96 (L96) model — a conceptual representation of geophysical turbulence:

$$\frac{du_i}{dt} = \left(-u_{i-1}\left(u_{i-2} - u_{i+1}\right) - u_i + f - \frac{hc}{J}\sum_{j=1}^{J} v_{i,j}\right) + \sigma_{u_i}\dot{W}_{u_i}, \quad i = 1, \ldots, I,$$

$$\frac{dv_{i,j}}{dt} = \left(-bcv_{i,j+1}\left(v_{i,j+2} - v_{i,j-1}\right) - cv_{i,j} + \frac{hc}{J}u_i\right) + \sigma_{v_{i,j}}\dot{W}_{v_{i,j}}, \quad j = 1, \ldots, J,$$

with $I = 40$ and $J = 4$. A weak scale separation is adopted to better mimics the real atmosphere with chaotic/turbulent behavior.





(from Wilks 2005)

Two approximate models.

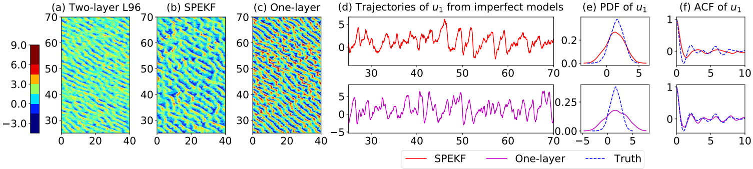
- One-layer L96 (L96-1LYR) model,

$$\frac{du_i}{dt} = \left(-u_{i-1}\left(u_{i-2} - u_{i+1}\right) - u_i + f\right) + \sigma_{u_i}\dot{W}_{u_i}, \quad i = 1 \dots, I.$$

- Stochastic parameterized imperfect model (L96-SP),

$$\frac{du_i}{dt} = \left(-u_{i-1}\left(u_{i-2} - u_{i+1}\right) - u_i + f - \frac{hc}{J}\sum_{j=1}^{J} v_{i,j}\right) + \sigma_{u_i}\dot{W}_{u_i}, \quad i = 1, \dots, I,$$

$$\frac{dv_{i,j}}{dt} = -\hat{d}_{i,j}(v_{i,j} - \hat{v}_{i,j}) + \hat{\sigma}_{v_{i,j}}\dot{W}_{v_{i,j}}, \quad j = 1, \dots, J.$$



Two approximate models.

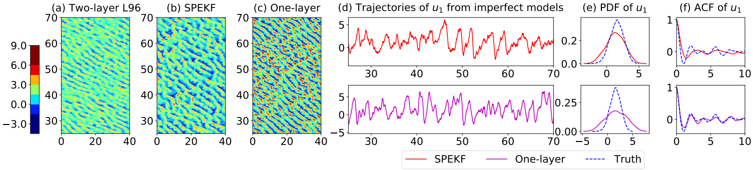
- One-layer L96 (L96-1LYR) model,

$$\frac{du_i}{dt} = \left(-u_{i-1}\left(u_{i-2} - u_{i+1}\right) - u_i + f\right) + \sigma_{u_i}\dot{W}_{u_i}, \quad i = 1. \ldots, I.$$

- Stochastic parameterized imperfect model (L96-SP),

$$\frac{du_i}{dt} = \left(-u_{i-1}\left(u_{i-2} - u_{i+1}\right) - u_i + f - \frac{hc}{J}\sum_{j=1}^{J} v_{i,j}\right) + \sigma_{u_i}\dot{W}_{u_i}, \quad i = 1, \ldots, I,$$

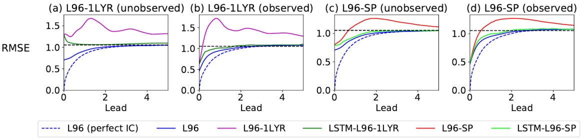
$$\frac{dv_{i,j}}{dt} = -\hat{d}_{i,j}(v_{i,j} - \hat{v}_{i,j}) + \hat{\sigma}_{v_{i,j}}\dot{W}_{v_{i,j}}, \quad j = 1, \ldots, J.$$



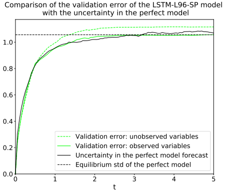
Setup of the experiments:

The observations are adopted only for the large-scale variables and are only on the even grid points: $u_2, u_4, \ldots, u_{40}$.

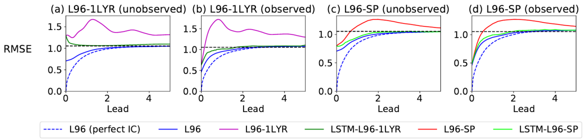
Long short-term memory (LSTM) NN models are trained based on each approximate model and are applied for forecast all the large-scale variables.

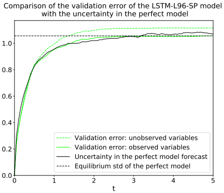


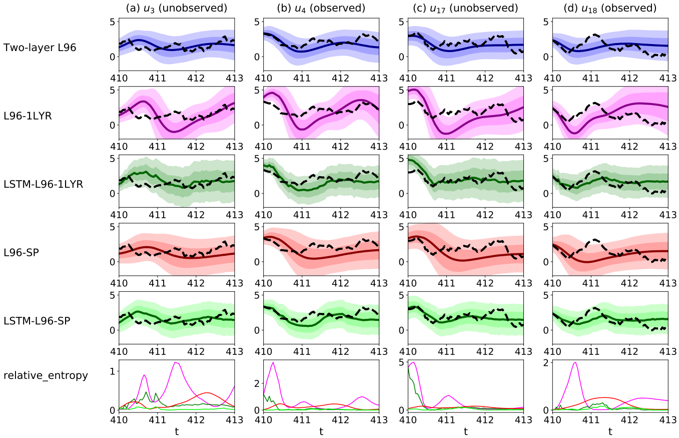
(a) L96-1LYR (unobserved) (b) L96-1LYR (observed) (c) L96-SP (unobserved) (d) L96-SP (observed)



Comparison of the validation error of the LSTM-L96-SP model with the uncertainty in the perfect model


(a) L96-1LYR (unobserved)
(b) L96-1LYR (observed)
(c) L96-SP (unobserved)
(d) L96-SP (observed)
RMSE
0.0
0.5
1.0
1.5
0
2
4
Lead
L96 (perfect IC)
L96
L96-1LYR
LSTM-L96-1LYR
L96-SP
LSTM-L96-SP

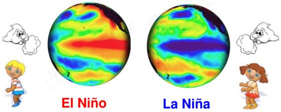

Comparison of the validation error of the LSTM-L96-SP model
with the uncertainty in the perfect model
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
t
Validation error: unobserved variables
Validation error: observed variables
Uncertainty in the perfect model forecast
Equilibrium std of the perfect model


(a) u_3 (unobserved)
(b) u_4 (observed)
(c) u_{17} (unobserved)
(d) u_{18} (observed)
Two-layer L96
L96-1LYR
LSTM-L96-1LYR
L96-SP
LSTM-L96-SP
relative_entropy
0
5
0
1
2
410
411
412
413
t

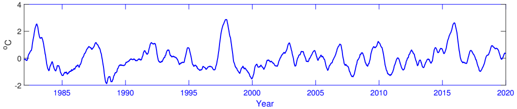
Application: Predicting a Widely Used El Niño-Southern Oscillation (ENSO) Index

(Chen, Harlim & Gilani, *GRL*, 2021)

ENSO is a large-scale interannual climate variability. It strongly connects with global warming and climate change.



The observations are short. The averaged sea surface temperature (SST) anomaly in the eastern Pacific (i.e., the Nino 3 index):



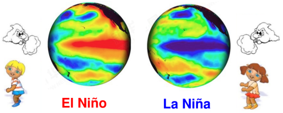
A suitable model describing the Nino 3 index:

$$
\begin{aligned}
\mathrm{d}T_E &= (-d_T T_E + \omega H_W + \alpha_T \tau)\,\mathrm{d}t + \sigma_T\,\mathrm{d}W_T,\\
\mathrm{d}H_W &= (-d_H H_W - \omega T_E + \alpha_H \tau)\,\mathrm{d}t + \sigma_H\,\mathrm{d}W_H,\\
\mathrm{d}\tau &= (-d_\tau \tau)\,\mathrm{d}t + \sigma_\tau(T_E)\,\mathrm{d}W_\tau.
\end{aligned}
$$

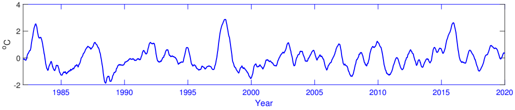
Application: Predicting a Widely Used El Niño-Southern Oscillation (ENSO) Index

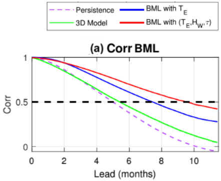
(Chen, Harlim & Gilani, *GRL*, 2021)

ENSO is a large-scale interannual climate variability. It strongly connects with global warming and climate change.



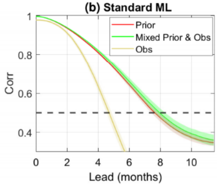
The observations are short. The averaged sea surface temperature (SST) anomaly in the eastern Pacific (i.e., the Nino 3 index):





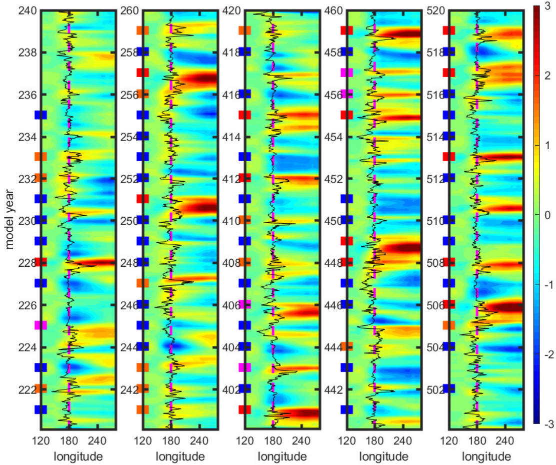
A suitable model describing the Nino 3 index:

$$\mathrm{d}T_E = (-d_T T_E + \omega H_W + \alpha_T \tau)\,\mathrm{d}t + \sigma_T\,\mathrm{d}W_T,$$
$$\mathrm{d}H_W = (-d_H H_W - \omega T_E + \alpha_H \tau)\,\mathrm{d}t + \sigma_H\,\mathrm{d}W_H,$$
$$\mathrm{d}\tau = (-d_\tau \tau)\,\mathrm{d}t + \sigma_\tau(T_E)\,\mathrm{d}W_\tau.$$

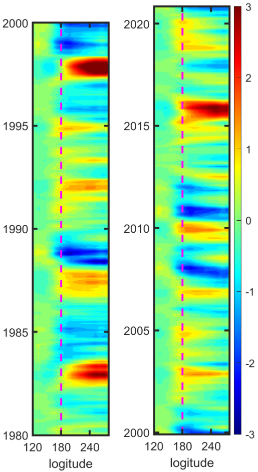


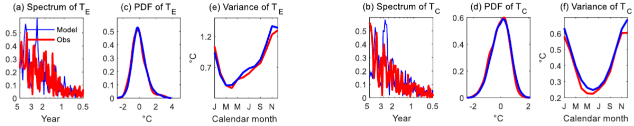
Ongoing ... ML prediction of the ENSO complexity using a recently developed stochastic model as the prior.

Hovmoller diagrams of the standard run



Observations (regressed)





Conclusion

- DA is important for uncertainty quantification in complex nonlinear systems.
- DA provides a new way to efficiently solve high-dimensional Fokker-Planck equations.
- DA connects parametric models with ML tools to improve the forecast.

Thank you!

(chennan@math.wisc.edu)