# ISDA-Online

**December 03, 2021 15 – 17 UTC**

# "Machine Learning for Data Assimilation"

*Organizers:* ***Nora Schenk*** *(DWD, Germany)*
***Marc Bocquet*** *(CEREA, École des Ponts and EDF R&D, France)*
***Manuel Pulido*** *(UMI-IFAECI (CNRS-CONICET-UBA) & Universidad Nacional del Nordeste, Argentina)*
***Lars Nerger*** *(AWI, Germany)*

**Program:**

**15:00 – 15:05** **Welcome**

**15:05 – 15:25** **VAE as a Stochastic Multidimensional Extension to Gaussian Anamorphosis**
Daisuke Hotta

**15:25 – 15:45** **State, Global and Local Parameter Estimation using Local Ensemble Kalman Filters: Applications to Online Machine Learning of Chaotic Dynamics**
Quentin Malartic, Alban Farchi, Marc Bocquet

**15:45 – 16:05** **Machine Learning Techniques to Construct Patched Analog Ensembles for Data Assimilation**
Lucia Minah Yang, Ian Grooms, Zofia Stanley

**16:05 – 16:10** **Time buffer**

**16:10 – 16:30** **Deep Learning for Retrieving Terrestrial Water Storage (TWS) from Spaceborne Gravity Observations and Satellite Altimetry**
Maria Aufschlager, Christopher Irrgang, Jan Saynisch-Wagner, Robert Dill, Eva Boergens, Maik Thomas

**16:30 – 16:50** **Forecast Uncertainty for Data Assimilation using Neural Networks**
Maximiliano A. Sacco, Yicun Zhen, Pierre Tandeo, Manuel Pulido, Juan Ruiz

**16:50 – 17:00** **Closing: Information on upcoming sessions**

Please note:

- When you login to the session before 15:00 UTC, and everything is quiet, this is most likely because we muted the microphones.
- The times in UTC are approximate. In case of technical problems, we might have to change the order of the presentations.
- **Time Zones**: 15 – 17 UTC

| | | |
|---|---|---|
| 03 – 05 pm GMT (London) | 04 – 06 pm CET (Berlin) | |
| 11 – 01 am CST (Shanghai) | 00 – 02 am JST (Tokyo) | 02 – 04 am AEDT (Sydney) |
| 07 – 09 am PST (San Fran.) | 08 – 10 am MST (Denver) | 10 – 12 am EST (New York) |

# VAE as a Stochastic Multidimensional Extension to Gaussian Anamorphosis

Daisuke Hotta[1]

[1]Meteorological Research Institute, Japan Meteorological Agency, Japan

Two-dimensional image data, such as satellite and radar imagery, are difficult to assimilate with the conventional data assimilation (DA) methods due to their (1) non-Gaussian error distribution, (2) dimensional redundancy, and (3) strong inter-pixel correlations.

Several techniques like Gaussian Anamorphosis have been shown to effectively alleviate the non-Gaussianity issue, but such methods have only been applied in a univariate manner due to techinical difficulty. The univariate nature of those methods makes it difficult to handle the issues (2) and (3).

Data compression through principal component analysis have been proposed to resolve the issue (2), and it is now becoming fairly standard to incorporate off-diagonal components in the observation error matrix R in operational DA systems to cope with the issue (3), but all these methods rely on Gaussianity assumption. To the author's knowledge, no single method hitherto proposed in DA literature appears to simultaneously handle all these issues.

In this study, we propose to use Variational Auto-Encoder (VAE) as a stochastic multidimensional extension to Gaussian Anamorphosis to simultaneously resolve the three difficulties listed above. VAE has been devised in Machine Learning literature as a generative model that learns, given a large enough dataset, how to stochastically convert complex multidimensional data to its latent variable in low-dimensional space (called decoder) and vice versa (called encoder). Thus, by using the VAE decoder, we can convert non-Gaussian high-dimensional data into a Gaussian latent space, perform regular Gaussian-based DA methods, then transform the assimilated results to the high-dimensional space by applying VAE encoder.

A preliminary assessment using a toy model that imitates assimilation of satellite infrared image to correct tropical cyclone position error shows promissing results. Several ideas toward more realistic DA problem will be also discussed in the presentation.

# State, Global and Local Parameter Estimation using Local Ensemble Kalman Filters: Applications to Online Machine Learning of Chaotic Dynamics

Quentin Malartic[1], Alban Farchi[1], Marc Bocquet[1]

[1]CEREA, École des Ponts and EDF R&D, France

In a recent methodological paper [2], we have shown how to learn chaotic dynamics along with the state trajectory from sequentially acquired observations, using local ensemble Kalman filters. Here, we more systematically investigate the possibilty to use a local ensemble Kalman filter with either covariance localization or local domains, in order to retrieve the state and a mix of key global and local parameters. Global parameters are meant to represent the surrogate dynamics, for instance through a neural network, which is reminiscent of data-driven machine learning of dynamics, while the local parameters typically stand for the forcings of the model. A family of algorithms for covariance and local domain localization is proposed in this joint state and parameter filter context. In particular, we show how to rigorously update global parameters using a local domain EnKF such as the LETKF, an inherently local method. The approach is tested with success on the 40-variable Lorenz model using several of the local EnKF flavors. A two-dimensional illustration based on a multi-layer Lorenz model is finally provided. It uses radiance-like non-local observations, and both local domains and covariance localization in order to learn the chaotic dynamics, and the local forcings. This presentation more generally addresses the key question of online machine learning with ensemble methods.

[1] - Quentin Malartic, Alban Farchi, Marc Bocquet. State, global and local parameter estimation using local ensemble kalman filters: applications to online machine learning of chaotic dynamics. SIAM/ASA Journal on Uncertainty Quantification, 0, 0. Submitted (2021).
[2] - Marc Bocquet, Alban Farchi, Quentin Malartic. Online learning of both state and dynamics using ensemble Kalman filters. Foundations of Data Science, 2021, 3 (3) : 305-330. doi: 10.3934/fods.2020015.

# Machine Learning Techniques to Construct Patched Analog Ensembles for Data Assimilation

Lucia Minah Yang[1], Ian Grooms[2], Zofia Stanley[3]

[1]CIMS, New York University, USA
[2]University of Colorado Boulder, USA
[3]CIRES Cooperative Institute for Research in Environmental Sciences, USA

We propose to use analogs of the forecast mean to generate an ensemble of perturbations for use in ensemble optimal interpolation (EnOI) or ensemble variational (EnVar) methods (see Grooms 2020 QJRMS). In addition to finding analogs from a library, we propose a new method of constructing analogs using autoencoders (a machine learning method). To extend the scalability of constructed analogs for use in data assimilation on geophysical models, we propose using patching schemes to divide the global spatial domain into digestable chunks. Using patches makes training the generative models possible and has the added benefit of being able to exploit parallel computing powers. The resulting analog methods using analogs from a catalog (AnEnOI), constructed analogs (cAnEnOI), and patched constructed analogs (p-cAnEnOI) are tested in the context of a multiscale Lorenz-‘96 model with O(1e3) variables, with standard EnOI and an ensemble square root filter (ESRF) for comparison.

The use of analogs from a modestly-sized catalog is shown to improve the performance of EnOI, with limited marginal improvements resulting from increases in the catalog size. The method using constructed analogs is found to perform as well as a full ESRF, and to be robust over a wide range of tuning parameters. We find that p-cAnENOI with larger patches produces the best data assimilation performance despite having larger reconstruction errors. All patch variants except for the variant that uses the smallest patch size outperform cAnEnOI as well as some traditional data assimilation methods such as the ensemble square root filter. Lastly, we show some results from applying cAnEnOI to Quasi-Geostrophic Coupled Model (q-gcm), a 2D model with O(1e7) variables.

# Deep Learning for Retrieving Terrestrial Water Storage (TWS) from Spaceborne Gravity Observations and Satellite Altimetry

Maria Aufschlager[1], Christopher Irrgang[1], Jan Saynisch-Wagner[1], Robert Dill[1], Eva Boergens[1], Maik Thomas[1]

[1]GFZ German Research Centre for Geosciences, Germany

Quantifying terrestrial water storage (TWS), an important component of the global water cycle, is crucial to understand its sensitivity to climate change and to predict resulting impacts on water resources, agriculture, and others. Numerical hydrology models and space-borne data, e.g., from the Gravity Recovery and Climate Experiment (GRACE) satellite mission, are used to study TWS variations. While TWS derived from GRACE is available for more than 20 years, its spatial resolution is limited to approximately 300km, making it impossible to show smaller scale structures like river networks. As a continuation of the published paper by Irrgang et al. (2020), a downscaling neural network is built, combining numerical modelling, space-borne observations and satellite altimetry. This combination of tools and data products is achieved through an adaptive validation term in the neural network loss function that mimics a non-linear data assimilation scheme. In this study, we show that the neural network retrieves TWS from synthetic GRACE observations with increased resolution from 0.5° to 0.125°. In validated regions, the neural network downscaling can outperform unconstrained forward simulations with respect to independent observations from satellite altimetry. Moreover, we were able to successfully apply the trained neural network on previously unseen simulated space-borne gravity fields, paving the way towards downscaling and using real GRACE observations.

Irrgang, C., Saynisch-Wagner, J., Dill, R., Boergens, E., & Thomas, M. (2020). Self-validating deep learning for recovering terrestrial water storage from gravity and altimetry measurements. Geophysical Research Letters, 47, e2020GL089258. https://doi.org/10.1029/2020GL089258

# Forecast Uncertainty for Data Assimilation using Neural Networks

Maximiliano A. Sacco[1], Yicun Zhen[2], Pierre Tandeo[2],
Manuel Pulido[3,4], Juan Ruiz[1,3,5,6]

[1]Universidad de Buenos Aires, Argentina
[2]IMT-Atlantique, France
[3]UMI-IFAECI (CNRS-CONICET-UBA), Argentina
[4]Universidad Nacional del Nordeste, Argentina
[5]CIMA Centro de Investigaciones del Mar y la Atmósfera, Argentina
[6]CONICET-UBA, Argentina

A fundamental aspect of data assimilation techniques is the quantification of forecast error uncertainty, as this has a major impact on the quality of the analysis produced and, consequently, on the forecasts generated from it. Most of the current operational assimilation systems obtain a state-dependent uncertainty quantification based on ensemble forecasts. However, this methodology is computationally expensive. In this work, we use a fully connected two-layer hidden neural network for the quantification of state-dependent forecast uncertainty in the context of data assimilation. The input to the network is a set of two consecutive forecast states, the initial condition and the desired time forecast of the analysis. The output of the network is a corrected forecast state and an estimate of its uncertainty.

Two methods are proposed. The first method consists of training a neural network using a loss function that estimates uncertainty in a local and semi-supervised manner. While in the second method we use a state space transformation prior to training in order to use a cost function based on the transformed observation probability. For training, we use a large database of forecasts and their corresponding analysis previously calculated. We performed simulation experiments of observation systems using the Lorenz'96 model as a proof of concept for an evaluation of the techniques, and compared it with classical ensemble-based approaches.

The results show that both proposed approaches can produce state-dependent estimates of forecast uncertainty without the need for an ensemble of states (at much lower computational cost), especially in the presence of model errors.