Learning ocean dynamical priors from noisy data using assimilation-derived neural nets

Said Ouala, Duong Nguyen, Cédric Herzet, Lucas Drumetz, Bertrand Chapron, Ananda Pascual, Fabrice Collard, Lucile Gaultier, Ronan Fablet

LEARNING OCEAN DYNAMICAL PRIORS FROM NOISY DATA USING ASSIMILATION-DERIVED NEURAL NETS

Said Ouala[1], *Duong Nguyen* [1], *Cédric Herzet*[1,2], *Lucas Drumetz*[1], *Bertrand Chapron*[3]
Ananda Pascual[4], *Fabrice Collard*[5], *Lucile Gaultier*[5], *Ronan Fablet*[1]

(1) IMT Atlantique; Lab-STICC, Brest, France
(2) INRIA Bretagne-Atlantique, Fluminance, Rennes, France
(3) Ifremer, LOPS, Brest, France
(4) IMEDEA, UIB-CSIC, Esporles, Spain
(5) ODL, Brest, France

ABSTRACT

Recent studies have investigated the identification of governing equations of geophysical systems from data. Here, we investigate such identification issues for ocean surface dynamcis from ocean remote sensing data. From a methodological point of view, we address the learning of data-driven dynamical models when only provided with a noisy training dataset. We propose a novel architecture that relies on data assimilation schemes to learn the underlying dynamical model through the minimization of a reconstruction cost. We demonstrate the relevance of the proposed architecture with respect to the state-of-the-art approaches in the identification and forecasting of synthetic and real case-studies.

***Index Terms*—** Dynamical systems, Data-driven models, Neural networks, Data assimilation

1. PROBLEM STATEMENT AND RELATED WORK

The constantly increasing wealth of simulation and observation data on geophysical dynamics makes data-driven strategies more and more appealing as new means to address key issues in ocean and atmosphere science, including for instance forecasting and assimilation issues [1]. In this respect, recent studies have investigated data-driven strategies to identify governing equations from data using different machine learning frameworks [2, 3].

The availability of representative training datasets is a strong requirement for the development of such approaches. When considering observation datasets (e.g., satellite-derived data or in-situ observations), the question of whether one may learn such data-driven representations from noisy and partial observation data naturally arises. For instance, regarding sea surface dynamics, beyond observation noise patterns, satellite sensors also involve irregular space-time sampling patterns due to their intrinsic characteristics or their sensitivity to the atmospheric conditions. In this work, we investigate these issues. We show that the effectiveness of previously proposed learning-based methods [2, 3, 4] is strongly affected when applied to noisy and partial observation datasets. Within a neural-network-based framework, we address the data-driven identification of governing equations as the joint learning of a dynamical model and the assimilation of the hidden states from noisy and partial observations. We restate state-of-the-art assimilation schemes as neural network architectures, such that the identification of the dynamical operator comes to the minimization of a data assimilation cost, rather than a simple forecasting error. From numerical experiments we demonstrate the relevance of the proposed approaches in the data-driven identification of dynamical operators.

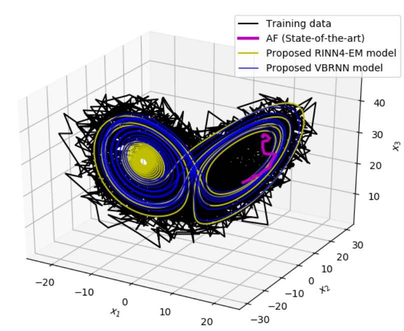

Fig. 1*: **Generated Lorenz-63 time series from data-driven models trained on noisy data.** Given the same initial condition, we generate time series of 1500 time steps for the following models : AF : Classic analog forecasting, RINN4-EM : proposed residual network optimized recursively on an assimilated training set, VBRNN : proposed variational bilinear recurrent network.*

This work was supported by GERONIMO project (ANR-13-JS03-0002), Labex Cominlabs (grant SEACS), Region Bretagne, CNES (grant OSTST-MANATEE), Microsoft (AI EU Ocean awards) and by MESR, FEDER, Région Bretagne, Conseil Général du Finistère, Brest Métropole and Institut Mines Télécom in the framework of the VIGISAT program managed by "Groupement Bretagne Télédétection" (BreTel).

2. PROPOSED MODELS

In this section, we describe two data-driven models inspired by a data assimilation framework. The first one is a direct implementation of the Ensemble Kalman Smoother (EnKS) in the optimization of an approximate dynamical model. The second model is inspired by recent neural networks implementations of classic smoothing schemes.

2.1. EM-based learning of dynamical models

Data-driven dynamical models identification techniques usually aim at finding the hidden governing equations of a given set of temporally evolving states. Formally, assuming that we are provided with a set of the states $\mathbf{x}_t$, where $t \in \{1, ..., T\}$ is the temporal sampling of our time series, the goal of a data-driven technique is to approximate the following model :

$$\mathbf{x}_{t+1} = \mathcal{F}(\mathbf{x}_t) + \eta_t \quad (1)$$

where $\mathcal{F}$ is the approximate dynamical system formulated as a parametric [3, 2] or a non parametric model [4] and η is a noise vector accounting for uncertainties.

State-of-the-art techniques [4, 2] usually suppose perfect knowledge of the state variables $\mathbf{x}$ or propose specific denoising solutions in the noisy case. On the other hand, regarding geosciences, we are almost never provided with the true states $\mathbf{x}_t$ and we are only given observations $\mathbf{y}_t$ related to our states through an observation model as follow :

$$\mathbf{y}_{t+1} = \mathcal{H}(\mathbf{x}_{t+1}, \Omega_{t+1}) + \epsilon_t \quad (2)$$

with $\mathcal{H}$ is an observation operator, Ω_{t+1} a mask accounting for missing data and ϵ is a random noise accounting for uncertainties in the model. Therefore, when only provided with partial and noisy observations, state-of-the-art architectures are most likely to fail or would need specific tuning depending on the irregularity present in the observations.

From another point of view, and supposing that the dynamical model $\mathcal{F}$ is known, we can approach the state variables $\mathbf{x}_t$ given the observations $\mathbf{y}_t$ using data assimilation schemes. In this case, equations (1) and (2) are used in Bayesian filtering schemes to approach the posterior density distribution of the states given the observations resulting in an estimations of the state variables. From this point of view, one can exploit data assimilation schemes as a natural framework to deal with partial and noisy data in data driven identification of dynamical operators. Formally, for a given architecture of the approximate dynamical model $\mathcal{F}$ we can restate the learning of the parameters of the dynamical model (which is usually carried through an optimization of a forecasting cost) as the optimization of a reconstruction cost over the hidden states $\mathbf{x}$ based on a data assimilation technique. The simplest way to do so is to alternate the optimization of our model and the computation of the hidden states $\mathbf{x}$ in an Expectation Maximization (EM) like algorithm as follow :

— **E-step** : Based on the approximate dynamical model $\mathcal{F}$, we compute an estimate of the posterior distribution of the states given the observations $p(\mathbf{x}_t|\mathbf{y}_{1:T})$, $\forall t \in \{1, ..., T\}$ using a smoothing data assimilation scheme;

— **M-step** : We learn the dynamical model given the posterior distribution $p(\mathbf{x}_t|\mathbf{y}_{1:T})$ using a maximum likelihood criterion.

2.2. Variational learning of dynamical operators

For many real-life problems, the observation operator $\mathcal{H}$ is highly nonlinear, or the signal to noise ratio is too small making the derivation of the posterior distribution $p(\mathbf{x}_t|\mathbf{y}_{1:T})$ intractable. This motivates the development of new filtering/smoothing schemes based on a variational approach where we introduce an approximate distribution $q(\mathbf{x}_t|\mathbf{y}_{1:T})$ of the true posterior $p(\mathbf{x}_t|\mathbf{y}_{1:T})$ parametrized by a neural network [5].

This technique has shown illustrious results for several tasks, including image, text and speech generation, noisy and irregularly-sampled time series modelling [5]. In this work, we adapt this idea to the problem of data driven identification of dynamical models. Specifically, the dynamical system $\mathcal{F}$, the observation operator $\mathcal{H}$ and the approximate posterior distribution $q(\mathbf{x}_t|\mathbf{y}_{1:T})$ are parametrized by deep neural networks as illustrated in Fig 2. However, instead of jointly learning the three models using the Evidence Lower Bound (ELBO) as the objective function like in [5], we alternately learn the dynamical model and the approximate posterior distribution as follow :

— **Initialization** : First, the approximate posterior network q is initialized to reconstruct the observations themselves. ($\mathbf{x}_t = \mathbf{y}_t$).

— **Pseudo E-M** : Given a series $\mathbf{x}^q_{1:T}$ reconstructed by the approximate posterior distribution, the dynamical model is trained to minimize the forecasting error of the generated state sequences. Given $\mathcal{F}$, the approximate posterior network q and the observation network $\mathcal{H}$ are then trained to minimized the distance between the new reconstructed observation sequence $\mathbf{y}^{\mathcal{F}}_{1:T}$ and the true observation sequence $\mathbf{y}_{1:T}$. This process is repeated until all the networks converge.

3. NUMERICAL EXPERIMENTS

In this section, we evaluate the proposed models in the forecasting and assimilation of dynamical systems governed by an unknown dynamical model when only provided with noisy training data.

3.1. Synthetic case study

The Lorenz 63 dynamical system is a 3-dimensional model governed by the following ODE :

$$\left\{\begin{array}{lcl}\frac{dx_{t,1}}{dt} & = & \sigma\left(x_{t,2}-x_{t,2}\right)\\ \frac{dx_{t,2}}{dt} & = & \rho x_{t,1}-x_{t,2}-x_{t,1}x_{t,3}\\ \frac{dx_{t,3}}{dt} & = & x_{t,1}x_{t,2}-\beta x_{t,3}\end{array}\right. \quad (3)$$

Under parameterization $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$, this system involves chaotic dynamics with two attractors.

We simulate Lorenz-63 state sequences using the LOSDA ODE solver [6] with an integration step of 0.01. We then add Gaussian noise with several variance levels N and evaluate the learnt models given the noisy training data.

Experimental setting : We perform a quantitative analysis of the forecasting and data assimilation performance of the proposed schemes with respect to state-of-the-art identification techniques :

— **Analog forecasting** [4] (AF) : This model applies a locally-linear regression operator estimated from nearest neighbors of the current state and their successors [4] ;

— **Sparse regression model** [2] (SR) : This model computes a sparse regression over an augmented states vector based on second order polynomial representations. The learnt dynamical model is then integrated to compute forecasts using the LOSDA ODE solver [6].

— **Four-block Bi-linear Residual Neural Network (Runge-kutta 4 setting)** (BRNN) : In this model, we reproduce a runge kutta 4 integration scheme of an approximate Ordinary Differential Equation (ODE) governing our states. The parametrization of this approximate model is set as in [3] and the model is trained to optimize the forecasting error ;

— **Four-block Bi-linear Residual Neural Network trained in an EM scheme** (BRNN-EM) : The same model as above trained in an EM scheme as proposed in section 2. The data assimilation technique used in the E-step was an EnKS and the training of the BRNN model in the M-step was carried through a Gaussian likelihood maximization ;

— **Variational Bi-linear Recurrent Neural Network** (VBRNN) : the dynamical model $\mathcal{F}$ is the BRNN described above, the observation operator $\mathcal{H}$ is an identity operator, the approximate posterior $q(\mathbf{x}_t|\mathbf{y}_{1:T})$ is modeled by a 2-layer bi-directional LSTM with the size of 9. To increase the capacity of this network, we also add an encoder to embed the observation $\mathbf{x}_t$ to the hidden space of the LSTM and a decoder to reconstruct $\mathbf{y}_t$, as shown in Fig. 2. Both the encoder and the decoder are modeled by a fully connected network, with one hidden layer of size of 7.

Forecasting and assimilation experiment : We report the mean forecasting performance, the largest Lyapunov exponent of the data-driven models and the reconstruction RMSE in Tab. 1. The proposed models achieve the best performances in terms of forecasting error. This is due to the training configuration of the dynamical models that rely on assimilated states rather then directly using the noisy observations as a training set. Regarding the long term characteristics of the proposed models, the VRBNN scheme is the only model to keep generating long term Lorenz time series (characterized by a Lyapunov exponent similar to the true Lorenz exponent) up to a noise level of $N = 7$. Finally, the assimilation performance also supports the relevance of the proposed models since only the neural networks based models achieve the best scores for the highest noise level.

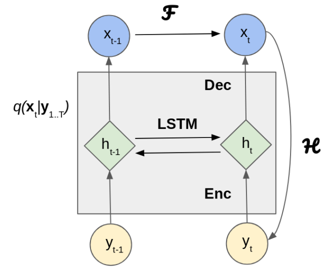

Fig. 2*: Architecture of the proposed VBRNN : x_t : true state at a given time step t, y_t : noisy observation, h_t : hidden state of the LSTM, $\mathcal{F}$ and $\mathcal{H}$ are the neural network implementation of the dynamical function and the observation operator, respectively. The backbone of the approximate posterior network $q(\mathbf{x}_t|\mathbf{y}_{1:T})$ is a LSTM, an encoder and a decoder network can be added to improve the modeling capacity of q.*

3.2. SLA case study

We also evaluate the relevance of the proposed EM based NN architecture for the forecasting and reconstruction of Sea Level Anomaly (SLA) dynamics [1].

We design an Observing System Simulation Experiment (OSSE) associated with the fast-sampling phase of the future mission assuming some noise patterns of the daily high-resolution SLA data. The data used in this experiment was generated by the WMOP model [8] with a $0.05°$ spatial resolution and a temporal resolution $h = 1$ day. We focus on the forecasting and reconstruction of fine scale components below 100km. Therefore, the large scale component was computed using an optimal interpolation and removed from the original data. As case-study region, we consider a region in the Algerian sea located on longitude $2.5°E$ to $4.25°E$ and latitude $36.5°S$ to $38.25°S$.

Our goal is to model and SLA anomaly time series using the proposed NN architecture. Following [9], project our data into an EOF subspace so that our patch is represented by a

1. We currently perform complementary experiments for this case-study using the proposed VBRNN approach that would be included in the final version.

Model		N=0.5	N=3	N=7
AF	$t_0 + h$	0.21	0.51	0.43
	$t_0 + 4h$	0.27	0.48	0.87
	λ_1	-1.251	-46.64	-29.404
	Assimilation	0.54	1.65	7.59
SR	$t_0 + h$	0.012	0.045	0.094
	$t_0 + 4h$	0.37	0.13	0.290
	λ_1	0.887	0.883	-0.240
	Assimilation	0.81	2.35	8.51
RINN4	$t_0 + h$	0.094	0.303	0.605
	$t_0 + 4h$	0.265	0.881	1.668
	λ_1	-0.574	-2.456	-2.241
	Assimilation	0.60	1.47	2.17
RINN4-EM	$t_0 + h$	0.001	0.011	0.018
	$t_0 + 4h$	0.008	0.025	0.035
	λ_1	0.904	0.042	-0.15
	Assimilation	0.57	1.39	2.10
VBRNN	$t_0 + h$	0.010	0.012	0.009
	$t_0 + 4h$	0.024	0.034	0.032
	λ_1	0.885	0.891	0.914
	Assimilation	0.57	1.42	2.13

Table 1*: **Forecasting and assimilation performance of data-driven models for Lorenz-63 dynamical model** : first two rows : mean RMSE for different forecasting time steps, third row : largest Lyapunov exponent of a predicted series of length of 10000 time-steps (The true largest Lyaponov exponent of the Lorenz 63 model is 0.90 [7]). Last row : data assimilation RMSE of a reconstructed a time series of length of 1000 time-steps when only observing the Lorenz states ones every 8 time-steps.*

15-dimensional vector. This EOF decomposition accounts for 92% of the total patch-level variance. Our training data was then corrupted with a Gaussian noise with several variance levels in the EOF space, which account for spatially-correlated noise patterns.

Forecasting and assimilation experiment : We report forecasting and assimilation performance in Tab.2. Similarly to the experiments with synthetic data, the proposed NN model trained in an EM scheme outperforms both the state-of-the-art techniques in terms of forecasting and assimilation score, with a relative gain up to 18% in terms of assimilation RMSE (the SR technique leads to a quasi-stationary short term forecasting models regardless of the noise level which explains the constant forecasting score).

4. CONCLUSION

In this work, we demonstrated the relevance of training data driven models in data assimilation schemes when provided with noisy training data. Further works could explore the relevance of such frameworks in the identification of sea surface dynamics based on real observations. The extension of the proposed framework to non-Gaussian systems through the use of a particle filter is also an interesting issue since it may extend the proposed framework to the representation of different distributions such as extremes.

Model		N=0.2	N=0.3	N=0.5
AF	$t_0 + h$	0.059	0.067	0.071
	$t_0 + 4h$	0.074	0.087	0.092
	Assimilation	0.041	0.050	0.062
SR	$t_0 + h$	0.017	0.017	0.017
	$t_0 + 4h$	0.043	0.044	0.044
	Assimilation	>0.5	>0.5	>0.5
RINN4	$t_0 + h$	0.016	0.031	0.037
	$t_0 + 4h$	0.050	0.055	0.078
	Assimilation	0.041	0.050	>0.5
RINN4-EM	$t_0 + h$	0.017	0.030	0.019
	$t_0 + 4h$	0.045	0.048	0.048
	Assimilation	0.042	0.045	0.051

Table 2*: **Forecasting and assimilation performance of data-driven models for the SLA case-study** : first two rows : mean RMSE for different forecasting time steps, third row : data assimilation RMSE of a reconstructed time series over the first 347 days of the year 2015. In the assimilation experiment, we used as observations synthetic along-track observations generated from real satellite track spatio-temporal locations from a four-altimeter sampling configuration.*

5. REFERENCES

[1] P. Tandeo, P. Ailliot, B. Chapron, R. Lguensat, and R. Fablet, "The analog data assimilation : application to 20 years of altimetric data," in *International Workshop on Climate Informatics*, Boulder, United States, sep 2015, pp. 1 – 2.

[2] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proc. Natl. Acad. Sci.*, vol. 113, no. 15, pp. 3932–3937, Apr. 2016.

[3] R. Fablet, S. Ouala, and C. Herzet, "Bilinear residual Neural Network for the identification and forecasting of dynamical systems," *SciRate*, dec 2017.

[4] R. Lguensat, P. Tandeo, P. Ailliot, M. Pulido, and R. Fablet, "The Analog Data Assimilation," *Mon. Weather Rev.*, aug 2017.

[5] M. Fraccaro, S. K. Sønderby, U. Paquet, and O. Winther, "Sequential Neural Models with Stochastic Layers," *arXiv :1605.07571 [cs, stat]*, may 2016, arXiv : 1605.07571.

[6] A. C. Hindmarsh, "ODEPACK, a systematized collection of ODE solvers," *IMACS Transactions on Scientific Computation*, vol. 1, pp. 55–64, 1983.

[7] J. C. Sprott, *Chaos and Time-Series Analysis*, Oxford University Press, Inc., New York, NY, USA, 2003.

[8] M. Juza, B. Mourre, L. Renault, S. Gómara, K. Sebastián, S. Lora, J. P. Beltran, B. Frontera, B. Garau, C. Troupin, M. Torner, E. Heslop, B. Casas, R. Escudier, G. Vizoso, and J. Tintoré, "Socib operational ocean forecasting system and multi-platform validation in the western mediterranean sea," *J. OPER. OCEANOGR.*, vol. 9, no. sup1, pp. s155–s166, 2016.

[9] R. Fablet, P. H. Viet, and R. Lguensat, "Data-Driven Models for the Spatio-Temporal Interpolation of Satellite-Derived SST Fields," *IEEE Trans. Comput. Imag.*, vol. 3, no. 4, pp. 647–657, dec 2017.