A Quantum Mechanical Approach for Data Assimilation in Climate Dynamics

Anonymous Authors[1]

Abstract

A framework for data assimilation in climate dynamics is presented, combining aspects of quantum mechanics, Koopman operator theory, and kernel methods for machine learning. This approach adapts the Dirac-von Neumann formalism of quantum dynamics and measurement to perform data assimilation (filtering) of climate dynamics, using the Koopman operator governing the evolution of observables as an analog of the Heisenberg operator in quantum mechanics, and a quantum mechanical density operator to represent the data assimilation state. The framework is implemented in a fully empirical, data-driven manner, using kernel methods for machine learning to represent the evolution and measurement operators via matrices in a basis learned from time-ordered observations. Applications to data assimilation of the Niño 3.4 index for the El Niño Southern Oscillation (ENSO) in a comprehensive climate model show promising results.

1. Introduction

Data assimilation is a framework for state estimation and prediction for partially observed dynamical systems (Majda & Harlim, 2012; Law et al., 2015). Adopting a predictor-corrector approach, it employs a forward model to evolve the probability distribution for the system state until a new observation is acquired, at which time that probability distribution is updated in an analysis step to a posterior distribution correcting for model error and/or uncertainty in the prior distribution. Since the seminal work of Kalman (1960) on filtering (which utilizes Bayes' theorem for the analysis step, under the assumption that all distributions are Gaussian), data assimilation has evolved to an indispensable tool in modeling and forecasting of the climate system.

Major challenges in data assimilation of the climate system are caused by the extremely large number of active degrees of freedom in the dynamics, evolving under nonlinear, and oftentimes partially known, equations of motion. These issues affect both the forward models (which invariably employ approximations such as subgrid-scale parameterization) and analysis procedures (where direct application of Bayes' theorem is generally not feasible). The past decades have seen vigorous research, leading to the creation of a vast array of operational data assimilation techniques for the weather and climate aiming to address these issues, including variational (Bannister, 2016), ensemble (Karspeck et al., 2018), and particle (van Leeuwen et al., 2019) methods.

In this work, we present a new method for data assimilation of climate dynamics combining aspects of quantum mechanics, operator-theoretic ergodic theory (Eisner et al., 2015), and kernel methods for machine learning. This approach results in a fully data-driven scheme, formulated entirely, and without approximation of the underlying dynamics, through linear evolution operators on spaces of observables. We demonstrate the utility of the method in a data assimilation experiment targeting the Niño 3.4 index for the El Niño Southern Oscillation (ENSO) in the Community Climate System Model Version 4 (CCSM4) (Gent et al., 2011).

2. Quantum mechanical data assimilation

The quantum mechanical data assimilation (QMDA) scheme employed in this work (Giannakis, 2019) is based on the observation that the predictor-corrector structure of data assimilation resembles the dynamics and measurement formalism of quantum mechanics, whereby the system state evolves under unitary dynamics between measurements (analogous to the forward modeling step in data assimilation) and under projective dynamics during measurements (the so-called "wavefunction collapse", analogous to the analysis step in data assimilation). Since quantum mechanics is inherently a statistical theory, formulated entirely through linear operators, it is natural to explore whether its formalism can be adapted to the context of data assimilation of classical dynamical systems. Importantly, such a scheme would be amenable to rigorous analysis and approximation through well-developed techniques in linear operator theory, harmonic analysis, and related fields.

In order to construct a concrete data assimilation scheme

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

from the abstract quantum mechanical axioms, we employ ideas from Koopman operator theory (Eisner et al., 2015). This framework characterizes a given measure-preserving dynamical system through intrinsically linear operators on a Hilbert space, called *Koopman operators* (Koopman, 1931; Kooopman & von Neumann, 1931), which govern the evolution of observables (functions of the state) under the dynamics. Mathematically, such Koopman operators have closely related properties to the Heisenberg evolution operators in quantum mechanics. Here, our starting point will be a continuous, measure-preserving, dynamical flow $\Phi^t : M \to M$, $t \in \mathbb{R}$, on a metric space M, with an ergodic, invariant, compactly supported Borel probability measure μ. In this setting M will play the role of the state space of the climate system (assumed inaccessible to direct observation), and μ the role of a climatological (equilibrium) distribution. The time-t Koopman operator associated with this system is the unitary operator $U^t : L^2(\mu) \to L^2(\mu)$ on the L^2 Hilbert space associated with the invariant measure, acting on vectors by composition with the dynamical flow, viz. $U^t f = f \circ \Phi^t$. We consider that the system is observed through a real-valued, bounded measurement function $h \in L^\infty(\mu)$.

With these definitions, we consider that the system evolves on some unknown dynamical trajectory $t \mapsto \Phi^t(x_0)$, starting from an arbitrary point $a_0 \in M$. The task of data assimilation is then to infer a probability distribution for the value of h at time $t > 0$, given the sequence of values $\tilde{h}_0, \tilde{h}_1, \dots, \tilde{h}_{J-1} \in \mathbb{R}$, with $\tilde{h}_n = h(a_n)$, $a_n = \Phi^{n\,\Delta t}(a_0)$, taken at times $t_n = n\,\Delta t$. Here, $\Delta t > 0$ is a fixed sampling interval. Following quantum mechanical ideas, we represent our knowledge about the system state by a time-dependent *density operator* $\rho_t : L^2(\mu) \to L^2(\mu)$, which can be loosely thought as a quantum mechanical analog of a probability measure in classical probability theory, and employ the quantum mechanical axioms in conjunction with the Koopman operator to evolve ρ_t in time. A probability measure $P_t : \mathcal{B}(\mathbb{R}) \to [0,1]$ on the Borel σ-algebra on $\mathbb{R}$ will then be derived from ρ_t, such that $P_t(\Omega)$ corresponds to the quantum mechanical probability that $h(\Phi^t(a_0))$ lies in a Borel set $\Omega \subseteq \mathbb{R}$. The full procedure, transcribing the standard Dirac-von Neumann axioms of quantum mechanics (Takhtajan, 2008), can be summarized as follows.

1. Associated with the data assimilation system is the Hilbert space $L^2(\mu)$, equipped with the inner product $\langle f, g\rangle_\mu = \int_M f^* g\, d\mu$. The state of the system lies in the set of non-negative, trace-class operators with unit trace. The observables of the data assimilation system are self-adjoint linear operators on $L^2(\mu)$. In particular, associated with the measurement function h is a self-adjoint, bounded multiplication operator T_h on $L^2(\mu)$, such that $T_h f = hf$.

2. Between measurements, i.e., for $t_n \le t < t_{n+1}$, the state evolves under the action of the unitary Koopman operators U^t induced by the dynamical flow. In particular, the state reached at time $t \in [t_n, t_{n+1})$ starting from a state ρ_{t_n} at time t_n is given by $\rho_t = U^{\tau *} \rho_t U^\tau$, where $\tau = t - t_n$.

3. As with every self-adjoint operator, T_h has a unique associated projection-valued measure $E_h : \mathcal{B}(\mathbb{R}) \to B(L^2(\mu))$, where $B(L^2(\mu))$ is the set of bounded linear operators on $L^2(\mu)$. That is, for every Borel set $\Omega \subseteq \mathbb{R}$, $E_h(\Omega)$ is an orthogonal projection operator on $L^2(\mu)$, and T_h can be decomposed through the spectral integral $T_h = \int_{\mathbb{R}} \omega\, dE_h(\omega)$. The set of values of h that can be observed with nonzero probability is given by the spectrum $\sigma(T_h)$, which coincides with the essential range of h.

4. If the data assimilation system has state ρ_t, then the probability that a measurement of h will yield a value lying in a Borel set $\Omega \subseteq \mathbb{R}$ is equal to $\operatorname{tr}(E_h(\Omega)\rho_t)$.

5. If the data assimilation state immediately before a measurement at time t_n is $\rho^-_{t_n}$, and a measurement of h yields the value $\omega \in \sigma(T_h)$, with $E_h(\{\omega\}) \neq 0$, then the state ρ_{t_n} immediately after the measurement is given by

$$\rho_{t_n} = E_h(\{\omega\})\rho^-_{t_n} E_h(\omega\})/ \operatorname{tr}(E_h(\{\omega\})\rho^-_{t_n} E_h(\{\omega\})).$$

3. Data-driven formulation

In a data-driven modeling scenario, we consider that available to us as training data is a time series $y_n = F(x_n)$ of the values of a continuous observation map $F : M \to \mathbb{R}^d$, sampled on a dynamical trajectory $x_n = \Phi^{n\,\Delta t}(x_0)$, $x_0 \in M$, with $0 \le n \le N-1$. For instance, in the ENSO application in section 4 below, the y_n will be snapshots of sea surface temperature (SST) sampled at d gridpoints on an Indo-Pacific domain. We also assume that the corresponding values $h_n = h(x_n)$ of the assimilated observable (in section 4, the Niño 3.4 index) are available. With these inputs, the data-driven formulation of QMDA scheme described in section 2 proceeds as follows.

First, following state-space reconstruction approaches (Sauer et al., 1991), we embed the snapshots y_n into a higher-dimensional space using the method of delays. Specifically, for a parameter $Q \in \mathbb{N}$, we construct the map $F_Q : M \to \mathbb{R}^{Qd}$, where $F_Q(x) = \big(F(x), F(\Phi^{\Delta t}(x)), \dots, F(\Phi^{(Q-1)\,\Delta t}(x))\big)$. Note that the values of F_Q can be empirically evaluated using the time-ordered training snapshots with out explicit knowledge of the underlying dynamical states on M, viz. $z_n := F_Q(x_n) = (y_n, y_{n+1}, \dots, y_{n-Q+1})$. This procedure is rigorously known to recover information lost about the underlying dynamical state if F is non-injective (which is the case in real-world applications). Therefore, all kernel calculations below will be performed on this extended data space (Giannakis & Majda, 2012; Berry et al., 2013).

Next, let $\mu_N = \sum_{n=0}^{N-1} \delta_{x_n}/N$ be the sampling probability measure on the training trajectory. Following Berry et al. (2015), we approximate the Koopman operator $U^{q\,\Delta t}$, $q \in \mathbb{Z}$ by the q-step *shift operator* $U_N^{(q)} : L^2(\mu_N) \to L^2(\mu_N)$ on the N-dimensional Hilbert space $L^2(\mu_N) \simeq \mathbb{C}^N$ associated with the sampling measure. The shift operator is then represented by an $L \times L$ matrix $\boldsymbol{U} = [U_{ij}]$, $U_{ij} = \langle \phi_i, U_N^{(q)} \phi_j \rangle_{\mu_N} = \sum_{n=0}^{N-q-1} \phi_i(x_n)\phi_j(x_{n+q})/N$, where the ϕ_i are orthonormal basis functions of $L^2(\mu_N)$ learned from the data z_n using kernel algorithms (Belkin & Niyogi, 2003; Coifman & Lafon, 2006; Berry & Harlim, 2016; Coifman & Hirn, 2013). Specifically, the ϕ_i are obtained as the leading eigenfunctions of a kernel integral operator $K : L^2(\mu_N) \to L^2(\mu_N)$, such that $Kf(x_m) = \sum_{n=0}^{N-1} k(z_m, z_n) f(x_n)/N$. Here, $k : \mathbb{R}^{Qd} \times \mathbb{R}^{Qd} \to \mathbb{R}_+$ is a continuous, symmetric positive-definite, Markov-normalized kernel function with exponential decay; see Giannakis (2019) for further details. Similarly, we approximate the density operator ρ_t and spectral projectors $E_h(\Omega)$ by finite-rank operators on $L^2(\mu_N)$, represented by $L \times L$ matrices in the $\{\phi_i\}$ basis. To approximate the spectral projectors, we construct a histogram of the training values h_n, whose bins, $\Omega_0, \dots, \Omega_{S-1} \subset \mathbb{R}$ have equal probability mass with respect to μ_N (i.e., equal number of training samples), and approximate $E_h(\Omega_i)$ by the multiplication operator on $L^2(\mu_N)$ by the characteristic function $\chi_{h^{-1}(\Omega_i)} : M \to \mathbb{R}$ of each bin. The corresponding measurement probabilities $P_t(\Omega_i) = \operatorname{tr}(\rho_t E_h(\Omega_i))$ can then be computed for any density operator ρ_t on $L^2(\mu_N)$.

4. Filtering the Niño 3.4 index in CCSM4

We now apply the framework described in sections 2 and 3 to perform data assimilation of the Niño 3.4 index in a 1300 yr control integration of CCSM4. Our experimental setup follows closely Slawinska & Giannakis (2017) and Giannakis & Slawinska (2018). As training data, we employ the first 1200 yr of monthly averaged SST fields on an Indo-Pacific domain, sampled at $d \approx 3 \times 10^4$ spatial points. We use time series of the Niño 3.4 index over the remaining 100 yr as the true signal $\tilde{h}(t) := h(\Phi^t(a_0))$ for data assimilation, where $h : M \to \mathbb{R}$ is the function returning the Niño 3.4 indices (in units of Kelvin) corresponding to the climate states in M. That signal is observed every 8 months in the data assimilation phase. Note that far more frequent observations would typically be available in an operational environment; here we work with a long observation interval to illustrate the skill of QMDA in predicting ENSO. We build the kernel k using $Q = 96$ delays (i.e., a physical embedding window of 8 yr), and use $L = 300$ corresponding eigenfunctions for operator approximation. We compute the measurement probabilities for $P_t(\Omega_i)$ for $S = 11$ equal-mass bins Ω_i.

Figure 1 shows the evolution of $P_t(\Omega_i)$ obtained via this approach over the first 30 yr of the data assimilation period,

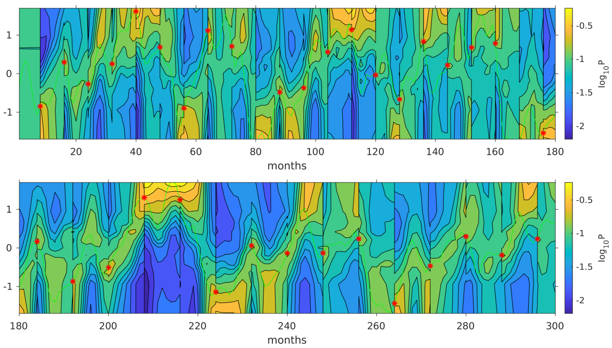

Figure 1. QMDA applied to the Niño 3.4 index in CCSM4. Contours indicate the logarithm of the time dependent measurement probabilities $P_t(\Omega_i)$ for the Niño 3.4 index to take values in bins Ω_i of equal probability mass $1/S = 1/11$ with respect to the invariant measure. The green line and red asterisks indicate the true signal $\tilde{h}(t)$ and observations $\tilde{h}_n = \tilde{h}(t_n)$, respectively.

in comparison to the true signal $\tilde{h}(t)$. In these experiments, we start from a purely uninformative density operator ρ_0, in the sense that the corresponding measurement probabilities $P_0(\Omega_i)$ are all equal to $1/S$. As soon as the first measurement is made at $t_1 = 8$ months, the forecast distribution collapses to a significantly more informative (peaked) distribution at $t = 1$ month. This distribution initially fails to track the true signal, but progressively improves over time as additional measurements are made. This is a manifestation of the fact that QMDA successfully assimilates the full history of measurements h_n to refine predictions of future ENSO states. Indeed, during the second half of the time interval depicted in Figure 1, $P_t(\Omega_i)$ tracks $\tilde{h}(t)$ with markedly higher precision and accuracy than during the first 15 years. It should be noted that this skill is achieved without QMDA being given any prior knowledge of the operating dynamics in the form of first-principles equations of motion or a statistical forecast model.

In conclusion, in this work we have demonstrated the potential of a new data assimilation framework for climate dynamics combining aspects of quantum mechanics, Koopman operator theory, and machine learning. Advantages of this framework include its fully nonparametric, "model-free" nature, and the fact that it is built entirely using finite-rank approximations of linear operators on Hilbert space, exhibiting rigorous convergence guarantees in the large data limit without relying on ad hoc approximations of the underlying dynamics and/or observation map. In addition, the framework naturally provides probabilistic output (as opposed to point forecasts), which is useful for uncertainty quantification, as well as risk assessment and decision making in operational scenarios. Besides ENSO, we expect QMDA to be useful in filtering a broad range of climate phenomena.

References

Bannister, R. N. A review of operational methods of variational and ensemble-variational data assimilation. *Quart. J. Roy. Meteorol. Soc.*, 143(703):607–633, 2016. doi: 10.1002/qj.2982.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15:1373–1396, 2003. doi: 10.1162/089976603321780317.

Berry, T. and Harlim, J. Variable bandwidth diffusion kernels. *Appl. Comput. Harmon. Anal.*, 40(1):68–96, 2016. doi: 10.1016/j.acha.2015.01.001.

Berry, T., Cressman, R., Gregurić-Ferenček, Z., and Sauer, T. Time-scale separation from diffusion-mapped delay coordinates. *SIAM J. Appl. Dyn. Sys.*, 12:618–649, 2013. doi: 10.1137/12088183x.

Berry, T., Giannakis, D., and Harlim, J. Nonparametric forecasting of low-dimensional dynamical systems. *Phys. Rev. E.*, 91:032915, 2015. doi: 10.1103/PhysRevE.91.032915.

Coifman, R. and Hirn, M. Bi-stochastic kernels via asymmetric affinity functions. *Appl. Comput. Harmon. Anal.*, 35(1):177–180, 2013. doi: 10.1016/j.acha.2013.01.001.

Coifman, R. R. and Lafon, S. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21:5–30, 2006. doi: 10.1016/j.acha.2006.04.006.

Eisner, T., Farkas, B., Haase, M., and Nagel, R. *Operator Theoretic Aspects of Ergodic Theory*, volume 272 of *Graduate Texts in Mathematics*. Springer, 2015.

Gent, P. R. et al. The Community Climate System Model version 4. *J. Climate*, 24:4973–4991, 2011. doi: 10.1175/2011jcli4083.1.

Giannakis, D. Quantum mechanics and data assimilation, 2019. URL https://arxiv.org/1903.00612.

Giannakis, D. and Majda, A. J. Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability. *Proc. Natl. Acad. Sci.*, 109(7):2222–2227, 2012. doi: 10.1073/pnas.1118984109.

Giannakis, D. and Slawinska, J. Indo-Pacific variability on seasonal to multidecadal time scales. Part II: Multiscale atmosphere-ocean linkages. *J. Climate*, 31(2):693–725, 2018. doi: 10.1175/JCLI-D-17-0031.1.

Kalman, R. E. A new approach to linear filtering and prediction problems. *J. Basic Eng.*, 82(1):35–45, 1960. doi: 10.1115/1.3662552.

Karspeck, A. R., Danabasoglu, G., J., A., Karol, S., Vertenstein, M., Raeder, K., Hoar, T., Neale, R., Edwards, J., and Craig, A. A global coupled ensemble data assimilation system using the Community Earth System Model and the Data Assimilation Research Testbed. *Quart. J. Roy. Meteor. Soc.*, 144(717):2404–2430, 2018. doi: 10.1002/qj.3308.

Kooopman, B. O. and von Neumann, J. Dynamical systems of continuous spectra. *Proc. Natl. Acad. Sci.*, 18(3):255–263, 1931. doi: 10.1073/pnas.18.3.255.

Koopman, B. O. Hamiltonian systems and transformation in Hilbert space. *Proc. Natl. Acad. Sci.*, 17(5):315–318, 1931. doi: 10.1073/pnas.17.5.315.

Law, K., Stuart, A., and Zygalakis, K. *Data Assimilation: A Mathematical Introduction*, volume 62 of *Texts in Applied Mathematics*. Springer, New York, 2015. doi: 10.1007/978-3-319-20325-6.

Majda, A. J. and Harlim, J. *Filtering Complex Turbulent Systems*. Cambridge University Press, Cambridge, 2012.

Sauer, T., Yorke, J. A., and Casdagli, M. Embedology. *J. Stat. Phys.*, 65(3–4):579–616, 1991. doi: 10.1007/bf01053745.

Slawinska, J. and Giannakis, D. Indo-Pacific variability on seasonal to multidecadal time scales. Part I: Intrinsic SST modes in models and observations. *J. Climate*, 30(14):5265–5294, 2017. doi: 10.1175/JCLI-D-16-0176.1.

Takhtajan, L. A. *Quantum Mechanics for Mathematicians*, volume 95 of *Graduate Series in Mathematics*. American Mathematical Society, Providence, 2008.

van Leeuwen, P. J., Künsch, H. R., Nerger, L., Potthast, R., and Reich, S. Particle filters for high-dimensional geoscience applications: A review. *Quart. J. Roy. Meteorol. Soc.*, 2019. doi: 10.1002/qj.3551.