Taming Uncertainty in a Complex World: The Rise of Uncertainty Quantification— A Tutorial for Beginners

Nan Chen, Stephen Wiggins, and Marios Andreou

George Box, a British statistician, wrote the famous aphorism, "All models are wrong, but some are useful." The aphorism acknowledges that models, regardless of whether they are qualitative, quantitative, dynamical, or statistical, always fall short of the complexities of reality. The ubiquitous imperfections of models come from various sources, including the lack of a perfect understanding of nature, limited spatiotemporal model resolutions due to computational power, inaccuracy in the initial and boundary conditions, etc. Therefore, uncertainty quantification (UQ), which quantitatively characterizes and estimates uncertainties, is essential to identify the usefulness of the model. UQ also defines the range of possible outcomes when certain aspects of the system are not precisely known. A key objective of UQ is to explore how uncertainties propagate, both through time evolution and across different quantities via complex nonlinear dependencies.

Characterizing Uncertainties

When uncertainty appears in the model and the input, the output can potentially take different values with typically unequal chances. Therefore, it is natural to characterize the output as a random variable, and the UQ of the model output can be based on the associated probability density function (PDF).

Nan Chen is an associate professor of mathematics at the University of Wisconsin-Madison. His email address is chennan@math.wisc.edu.

Stephen Wiggins is the William R. Davis '68 Distinguished Chair in Mathematics at the United States Naval Academy and a professor of applied mathematics at the University of Bristol. His email address is s.wiggins@bristol.ac.uk.

Marios Andreou is a PhD student of mathematics at the University of Wisconsin-Madison. His email address is mandreou@math.wisc.edu.

Intuitively, the spread of a distribution, which describes how close the possible values of the output are to each other, measures the uncertainty of a variable. If a PDF is Gaussian, variance is a natural indicator describing the spread. Therefore, the uncertainty associated with the three Gaussian distributions is expected to increase from Panel (a) to Panel (c) in Figure 1. On the other hand, non-Gaussian distributions are widely seen in practice due to the intrinsic nonlinearity of the underlying system. The non-Gaussian features should also be highlighted in the rigorous quantification of the uncertainty.

Shannon's entropy: Quantifying the uncertainty in one PDF. Denote by $p(x)$ the PDF of a random variable X. Shannon's entropy, an information measurement, is a natural choice to rigorously quantify uncertainty. It is defined as [Gra11]:

$$\mathcal{S}(p) = -\int p(x)\ln(p(x))\,\mathrm{d}x. \qquad (1)$$

Shannon's entropy originates from the theory of communication when a "word" is represented as a sequence of binary digits with length n, so the set of all words of length n has $2^n = N$ elements. Therefore, the amount of information needed to characterize one element, which is the number of digits, is $n = \log_2 N$. Consider the case where the entire set is divided into disjoint subsets, each with N_i total elements. The chance of randomly taking one element that belongs to the i-th subset is $p_i = N_i/N$. If an element belongs to the i-th subset, then the additional information to determine it is $\log_2 N_i$. Therefore, the average amount of information to determine an element is

$$\sum_i \frac{N_i}{N}\log_2 N_i = \sum_i \frac{N_i}{N}\log_2\left(\frac{N_i}{N}\cdot N\right)$$
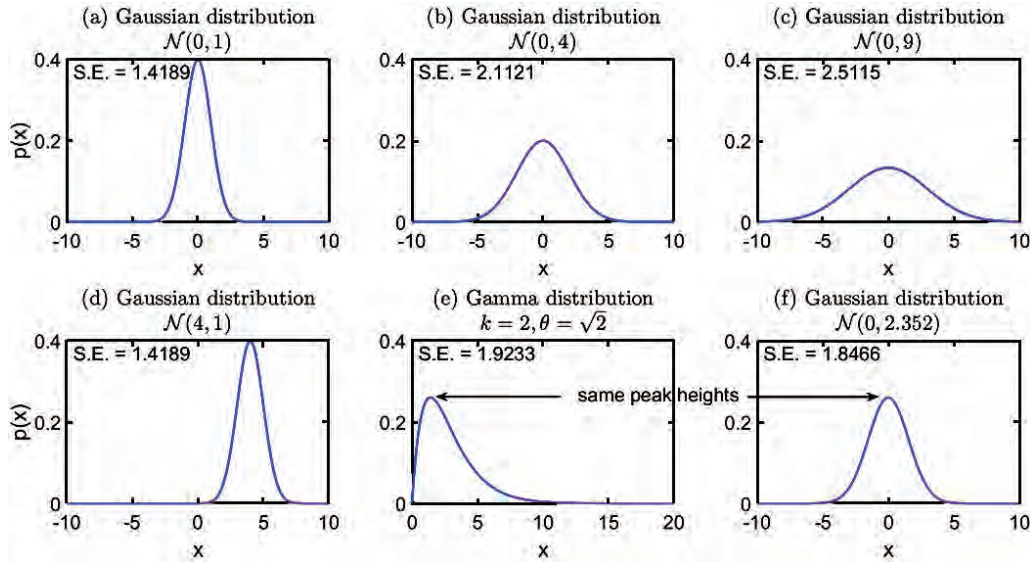$$= \sum_i p_i \log_2 p_i + \log_2 N. \qquad (2)$$

Figure 1. Quantifying the uncertainty using Shannon's entropy (S.E.) for Gaussian and non-Gaussian distributions. In all panels, the x-axis spans 20 units.

Recall that $\log_2 N$ is the information to determine an element given the full set. Thus, the corresponding average lack of information is $-\sum p_i \log_2 p_i$, which is the uncertainty. The formal definition of Shannon's entropy (1) generalizes the above argument. It exploits $-\ln(p(x))$, the negative of a natural logarithm function, to characterize the lack of information of each event x and then takes the continuous limit to replace the finite summation with an integral that represents the uncertainty averaged over all events. See [Sha48] for a more rigorous derivation of Shannon's entropy and its uniqueness. Notably, Shannon's entropy (1) applies to general non-Gaussian PDFs.

For certain distributions, $\mathcal{S}(p)$ can be written down explicitly. If $p \sim \mathcal{N}(\mu, R)$ is an m-dimension Gaussian distribution, where μ and R are the mean and covariance, then Shannon entropy has the following form (where "det" denotes the matrix determinant):

$$\mathcal{S}(p) = \frac{m}{2}(1 + \ln 2\pi) + \frac{1}{2}\ln \det(R). \qquad (3)$$

In the one-dimensional situation, (3) implies that the uncertainty is uniquely determined by the covariance and is independent of the mean value, consistent with the intuition. Therefore, Shannon's entropy confirms that the uncertainty increases from Panel (a) to Panel (c) in Figure 1, and the uncertainty has the same value in Panels (a) and (d). However, without a systematic information-based measurement, there is typically no single empirical indicator, like variance (or equivalently, peak height), that can fully characterize the uncertainty in complex non-Gaussian PDFs. Therefore, Shannon's entropy provides a systematic way for UQ. Panel (e) shows a Gamma distribution, which is skewed and has a one-sided fat tail. The fat tail usually corresponds to extreme events, which are

farther from the mean value, naturally increasing the uncertainty. Therefore, despite the same peak height of the two PDFs in Panels (e) and (f), the fat-tailed Gamma PDF has a larger entropy.

Relative entropy: Measuring how one PDF is different from another. In many practical problems, the interest lies in estimating the lack of information in one distribution $p^M(x)$ related to another $p(x)$. Typically, $p(x)$ is associated with a full probabilistic model while $p^M(x)$ comes from a reduced-order approximate model. The latter is often less informative but is widely used in practice to accelerate computations.

Recall that the uncertainty about x in $p(x)$ is $\mathcal{I}(p(x)) = -\ln(p(x))$. The average lack of information, i.e., Shannon's entropy, is

$$\langle \mathcal{I}(p(x)) \rangle = -\int p(x)\ln(p(x))\,\mathrm{d}x. \qquad (4)$$

However, while the uncertainty about x in $p^M(x)$ is $\mathcal{I}(p^M(x)) = -\ln(p^M(x))$, the expected lack of information under $p^M(x)$ is

$$\langle \mathcal{I}(p^M(x)) \rangle = -\int p(x)\ln(p^M(x))\,\mathrm{d}x. \qquad (5)$$

This is because even though the approximate model is used to measure the information $\ln(p^M(x))$, the actual probability of x to appear always comes from the full system, namely $p(x)$, which is objective and independent of the choice of the model. In other words, the role of the approximate model is to provide the lack of information for each event x. In contrast, the underlying distribution of the occurrence of x is objective regardless of the approximate model used. Relative entropy, also known
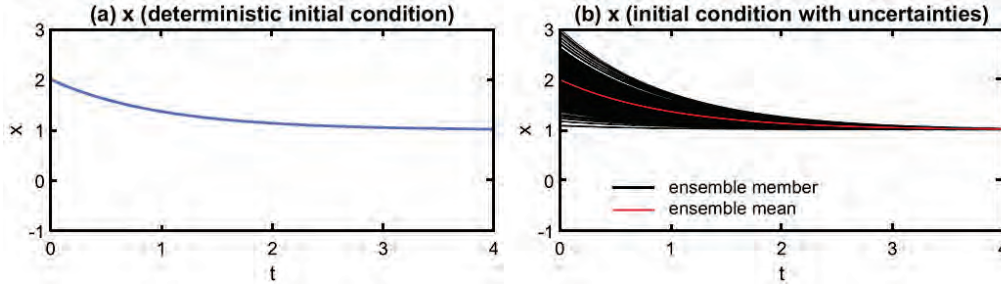
Figure 2. Solutions of the linear system (8). Panel (a): time evolution of $x(t)$ with a deterministic initial condition. Panel (b): time evolution of $x(t)$ with the initial condition given by a Gaussian distribution. The ensemble size is 1000.

as the Kullback-Leibler (KL) divergence, characterizes the difference between these two entropies [CT06]:

$$
\begin{aligned}
\mathcal{P}(p, p^M) &= \langle \mathcal{I}(p^M(x)) \rangle - \langle \mathcal{I}(p(x)) \rangle \\
&= \int p(x) \ln \left(\frac{p(x)}{p^M(x)} \right) dx.
\end{aligned}
\tag{6}
$$

The relative entropy $\mathcal{P}(p, p^M)$ is nonnegative. It becomes larger when p and p^M are more distinct. Since $p(x)$ appears in both $\langle \mathcal{I}(p^M(x)) \rangle$ and $\langle \mathcal{I}(p(x)) \rangle$, the relative entropy is not symmetric. One desirable feature of $\mathcal{P}(p, p^M)$ is that it is invariant under general nonlinear change of variables. As a remark, if $p^M(x) \ln(p^M(x))$ is utilized in (5) to compute the expected lack of information under $p^M(x)$, then the analog to (6) is called Shannon entropy difference.

To illustrate that relative entropy (6) is a more appropriate definition of the lack of information than Shannon entropy difference, consider a model to bet on a soccer game: Team A vs. Team B. A comprehensive model p gives the odds for Team A: 10% (win), 10% (draw), and 80% (lose). However, someone unfamiliar with soccer may have a biased model p^M, which gives the odds for Team A: 80% (win), 10% (draw), and 10% (lose). If entropy difference is used, the resulting lack of information will be precisely zero. In contrast, relative entropy considers the lack of information in p^M related to p regarding each outcome (win/draw/lose).

When both $p \sim \mathcal{N}(\mu, R)$ and $p^M \sim \mathcal{N}(\mu^M, R^M)$ are m-dimensional Gaussians, the relative entropy has the following explicit formula [MKC02],

$$
\begin{aligned}
\mathcal{P}(p, q) &= \frac{1}{2} \left[(\mu - \mu^M)^{\mathrm{T}} (R^M)^{-1} (\mu - \mu^M) \right] \\
&\quad + \frac{1}{2} \left[\mathrm{tr}(R(R^M)^{-1}) - m - \ln \det(R(R^M)^{-1}) \right],
\end{aligned}
\tag{7}
$$

where "tr" is the trace of a matrix. The first term on the right-hand side of (7) is called "signal," which measures the lack of information in the mean weighted by model covariance. The second term involving the covariance ratio is called "dispersion."

UQ in Dynamical Systems

From now on, UQ will be discussed in the context of dynamical systems. The uncertainty propagates in different ways in linear and nonlinear systems.

Examples of uncertainty propagation in linear and nonlinear dynamical systems. Consider a linear ordinary differential equation,

$$
\frac{dx}{dt} = -ax + f,
\tag{8}
$$

where $a > 0$ is the damping coefficient and f is an external forcing. The solution of (8) can be written down explicitly $x(t) = x(0)e^{-at} + (1 - e^{-at})f/a$, where $x(t)$ converges to f/a in an exponential rate. Consider the time evolution of the linear dynamics (8) with two different initial conditions. A deterministic initial condition $x(0) = 2$ is given in the first case. The time evolution of $x(t)$ is shown in Panel (a) of Figure 2. In the second case, uncertainty appears in the initial condition, which is given by a Gaussian distribution $x(0) \sim \mathcal{N}(2, 0.09)$. Different initial values are drawn from this distribution and follow the governing equation (8). The black curves in Panel (b) show the time evolution of different ensemble members, while the red curve is the ensemble average. The uncertainty dissipated over time, and the time evolution of the ensemble average follows the same trajectory as the deterministic case. In other words, the uncertainty does not change the mean dynamics.

Next, consider the Lorenz 63 model, which is a nonlinear chaotic system [Lor63],

$$
\frac{dx}{dt} = \sigma(y - x), \quad \frac{dy}{dt} = x(\rho - z) - y, \quad \frac{dz}{dt} = xy - \beta z.
\tag{9}
$$

Again, consider the time evolution of the solution starting from two sets of initial conditions, one deterministic $x(0) = 20, y(0) = -20, z(0) = 25$ and one containing uncertainty $x(0) \sim \mathcal{N}(20, 1), y(0) \sim \mathcal{N}(-20, 1), z(0) \sim \mathcal{N}(25, 1)$. By taking the standard parameter values $\sigma = 10$, $\beta = 8/3$ and $\rho = 28$, the system displays a chaotic behavior. Panel (a) of Figure 3 shows the attractor of the system, which resembles a butterfly. Panel (b) shows the time evolution of $z(t)$ starting from the deterministic initial
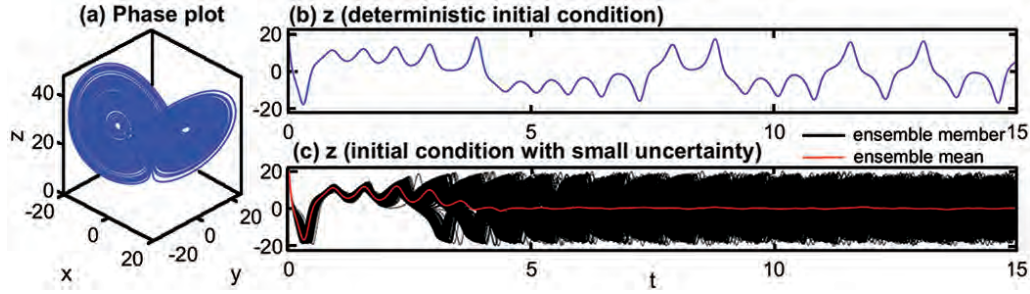
Figure 3. Solutions of the nonlinear chaotic Lorenz 63 model (9). Panel (a): the Lorenz attractor. Panel (b): time evolution of $z(t)$ with a deterministic initial condition. Panel (b): time evolution of $z(t)$ with the initial condition given by a Gaussian distribution. The ensemble size is 1000.

condition. In contrast, the black curves in Panel (c) show ensemble members starting from different values drawn from the given initial distribution, while the red curve is the ensemble average. Although the ensemble mean follows the trajectory in the deterministic case and has a small uncertainty within the first few units, the ensemble members diverge quickly. Notably, the ensemble average significantly differs from any model trajectories. This implies that uncertainty has a large impact on the mean dynamics. In other words, the mean evolution and the uncertainty cannot be considered separately, as in the linear case.

Impact of nonlinearity on uncertainty propagation. To understand the impact of nonlinearity on uncertainty propagation, let us decompose a random variable x into its mean and fluctuation parts via the Reynolds decomposition [Mül06]

$$x = \langle x \rangle + x', \qquad \text{with } \langle x' \rangle = 0, \tag{10}$$

where $\langle \cdot \rangle$ represents the ensemble average computed by taking the summation of all ensemble members and then dividing by the total number of ensembles. So, $\langle (x')^2 \rangle$ is the variance of x. For the linear system (2), taking the ensemble average leads to the mean dynamics

$$\frac{d\langle x \rangle}{dt} = (-a\langle x \rangle + f), \tag{11}$$

which implies the time evolution of the mean $\langle x \rangle$ is not affected by the uncertainty described by the fluctuation part. However, the situation becomes very different if nonlinear terms appear in the dynamics. Consider a quadratic nonlinear term on the right-hand side of the dynamics,

$$\frac{dx}{dt} = bx^2 + \dots. \tag{12}$$

Since

$$\langle x^2 \rangle = \langle (\langle x \rangle + x')^2 \rangle = \langle (\langle x \rangle^2 + (x')^2 + 2\langle x \rangle x') \rangle$$

$$= \langle x \rangle^2 + \langle (x')^2 \rangle, \tag{13}$$

taking the ensemble average of (12) leads to

$$\frac{d\langle x \rangle}{dt} = b\langle x \rangle^2 + b\langle (x')^2 \rangle + \dots, \tag{14}$$

which indicates that the higher-order moments containing the information of uncertainties affect the time evolution of the lower-order moments (e.g., the mean dynamics) via nonlinearity. It also reveals that the moment equations for general nonlinear dynamics will never be closed. In practice, approximations are made to handle the terms involving higher-order moments in the governing equations of the lower-order moments to form a solvable closed system [MQ18, Che23].

In chaotic systems, small uncertainties are quickly amplified by positive Lyapunov exponents, making an accurate state forecast/estimation challenging. Additional resources can be combined with models to facilitate uncertainty reduction in state estimation.

Uncertainty Reduction Via Data Assimilation (DA)

Model and observational data are widely utilized to solve practical problems. However, neither model nor observation is close to perfect in most applications. Models are typically chaotic and involve large uncertainties. Observations contain noise and are often sparse, incomplete, and indirect. Nevertheless, when a numerical model and observations are optimally combined, the estimation of the state can be significantly improved. This is known as data assimilation (DA). DA was initially developed in numerical weather prediction, which improves the initialization for a numerical forecast model. Since then, DA has become an essential tool for many applications, including dynamical interpolation of missing data, inferring the unobserved variables, parameter estimation, assisting control and machine learning, etc.

Denote by u the state variable and $v = gu + \epsilon$ the noisy observation with g the observational operator and ϵ a random noise, the underlying principle of DA is given by Bayes' theorem [Jef73]:

$$\underbrace{p(u|v)}_{\text{posterior}} \propto \underbrace{p(u)}_{\text{prior}} \times \underbrace{p(v|u)}_{\text{likelihood}}, \tag{15}$$

where "$\propto$" stands for "proportional to." In (15), $p(u)$ is the forecast distribution by using a model built upon prior
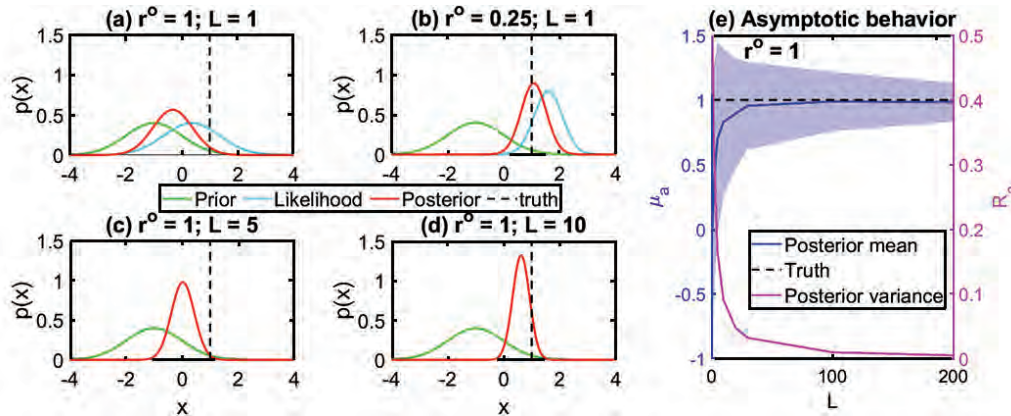
Figure 4. Posterior distribution from Bayes' formula (15) with $\mu_f = 1$, $R_f = 1$, and $m = 1$. Panels (a)–(d): the PDFs with different observational uncertainty r^o and number of observations L. To keep the figure concise, the PDFs corresponding to the $L = 5$ or 10 noisy observations are omitted from Panels (c)–(d). Panel (e): the asymptotic behavior of the posterior mean μ_a and the posterior variance R_a as a function of L. Due to the randomness in observations, the shaded area shows the variation of the results with 100 sets of independent observations for each fixed L.

knowledge, while $p(v|u)$ is the probability of observation under the model assumption. Their combination is the conditional distribution of u given v, which is called the posterior distribution. Notably, due to the additional information from observation, uncertainty is expected to be reduced from the prior to the posterior distribution.

Assume the prior distribution $p(u) \sim \mathcal{N}(\mu_f, R_f)$ and the observational noise with variance r^o are both Gaussian. Plugging these into (15), it is straightforward to show that the posterior distribution $p(u|v) \sim \mathcal{N}(\mu_a, R_a)$ is also Gaussian. The posterior mean μ_a and covariance R_a can be written down explicitly as

$$\mu_a = (I - Kg)\mu_f + Kv, \quad R_a = (I - Kg)R_f, \quad (16)$$

where I is an identity matrix of size $m \times m$ with m being the dimension of u and K is given by $K = R_f g^{\mathrm{T}}(gR_f g^{\mathrm{T}} + r^o)^{-1}$. Now consider the case with $m = 1$ and $g = 1$. Then, all three distributions in (15) are one-dimensional Gaussians, and the observation equals the truth plus noise. In such a case, $K = R_f/(R_f + r^o) \in [0, 1]$ and μ_a becomes a weighted summation of the prior mean and the observation with weights being $1 - K$ and K, respectively. When the observational noise is much smaller than the prior uncertainty, i.e., $r^o \ll R_f$, the posterior mean almost fully trusts the observation since $K \approx 1$. In contrast, if the observation is highly polluted, i.e., $r^o \gg R_f$, then $K \approx 0$ and the observational information can almost be ignored, and the posterior mean nearly equals the prior mean. Essentially, the weights are fully determined by the uncertainties. On the other hand, the posterior variance R_a in (16) is always no bigger than the prior variance R_f, indicating that the observation helps reduce the uncertainty. Panels (a)–(b) of Figure 4 validate the above conclusions. In some applications, repeated measurements with independent noises are available, which further advance the reduction of the posterior

uncertainty. For example, in the above scalar state variable case, when L repeated observations are used, the observational operator becomes a $L \times 1$ vector $g = (1, \dots, 1)^{\mathrm{T}}$ and similarly for the noise $\epsilon = (\epsilon_1, \dots, \epsilon_L)^{\mathrm{T}}$. According to Panels (a), (c), and (d), the posterior uncertainty reduces when the number of observations increases from $L = 1$ to 10. Panel (e) shows that when L becomes large, the posterior mean converges to the truth, and the posterior variance decreases to zero. Correspondingly, when characterizing the uncertainty reduction in the posterior distribution related to the prior via the relative entropy (7), the signal part converges to a constant while the dispersion part scales as $\ln(L)$. See the online supplementary document for more details.

When the forecast model and the observational operator are linear, and the noises are Gaussian, the above procedure is called the Kalman filter [Kal60]. Closed analytic formulae are available for the posterior distribution. However, models are generally nonlinear, and the associated PDFs are often non-Gaussian. Numerical methods, such as ensemble Kalman filter (EnKF) or particle filter, are widely used to find approximate solutions [LSZ15, ABN16].

In the following, UQ will be discussed in the context of Lagrangian DA (LaDA) [AJS08], which exploits Lagrangian observations. The same procedure for quantifying and reducing the uncertainty can be applied using other observations.

Lagrangian data assimilation (LaDA). Lagrangian tracers are moving drifters, such as robotic instruments, balloons, sea ice floes, and even litter. They are often used to recover the flow field that drives their motions. However, recovering the entire flow field based solely on Lagrangian tracers is challenging. This is because tracers are usually sparse, which prevents a direct estimation of the flow velocity in
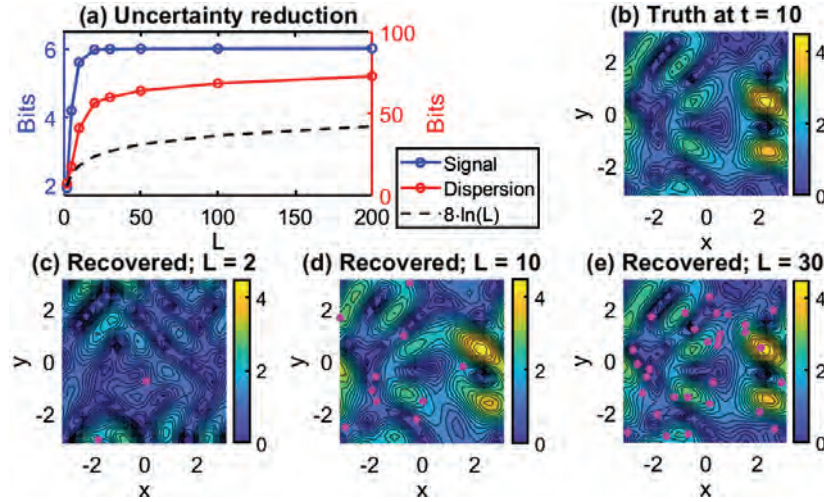
Figure 5. UQ in LaDA. Panel (a): uncertainty reduction in the signal and dispersion parts as a function of L. Panels (b)–(e): comparing the true flow field with the recovered ones using 2, 10, and 30 tracers.

regions with no observations. In addition, the measured quantity is the tracer displacements. Observational noise can propagate through the time derivative from displacement to velocity. A flow model, despite being generally turbulent, can provide prior knowledge about the possible range of velocity in the entire domain. Observational information dominates the state estimation at the locations covered by tracers. They also serve as the constraints, conditioned on which the uncertainty in estimating the flow field through the model in regions without observations can be significantly reduced. LaDA is widely used to provide a more accurate recovery of the flow field, facilitating the study of flow properties. For simplicity, assume the tracer velocities are equal to the underlying flow field. The LaDA scheme consists of two sets of equations: one for the observational process and the other for the flow model, described as follows:

$$\text{Obs process:} \quad \frac{\mathrm{d}x_l}{\mathrm{d}t} = u(x_l, t) + \sigma_x \dot{W}_l, \quad l = 1, \dots, L,$$

$$\text{Flow model:} \quad \frac{\partial u}{\partial t} = \mathcal{F}(u, t), \qquad (17)$$

where L is the number of tracers, and $\dot{W}_l$ is a white noise, representing the observational noise. Each x_l is a two-dimensional vector containing the displacements of the l-th tracer, and u is the two-dimensional velocity field (e.g., the surface of the ocean). Notably, the flow velocity in the observational process is typically a highly nonlinear function of displacement, making the LaDA a challenging nonlinear problem.

UQ in LaDA. One important UQ topic in LaDA is quantifying the uncertainty reduction in the estimated flow field as a function of the number of tracers L. The comparison is between the posterior distribution from LaDA and the prior one. The latter is the statistical equilibrium solution (i.e., the attractor) of the chaotic flow model. It is the best

inference of the flow field in the absence of observations. With two distributions involved in the comparison, the relative entropy (6) is a natural choice to measure the uncertainty reduction.

Under certain conditions [CMT14], closed analytic formulae are available for finding the posterior distribution of the nonlinear LaDA problem, which allows for using rigorous analysis to quantify the uncertainty reduction for the long-term behavior. By applying (7), it can be shown that the signal part of the relative entropy converges to a constant as $L \to \infty$, where the posterior mean μ in (7) is replaced by the true flow field. This is intuitive as the tracers become dense in the domain, and the observational error cancels off at each location when $L \to \infty$. On the other hand, the dispersion part in (7) will never converge. Instead, it grows at $\ln(L)$. Although it confirms that deploying extra tracers will always bring additional information, the increment of the uncertainty reduction will decrease as L becomes large. Specifically, the logarithm dependence on L implies that reducing the uncertainty by a fixed amount requires an exponential increase in the number of tracers. This means, in practice, once the information gain reaches a certain level, there is no need to continue deploying tracers as the uncertainty reduction from the additional tracers becomes marginal. Figure 5 shows a numerical simulation that validates the theoretic conclusion.

The Role of Uncertainty in Diagnostics

Parameter estimation. Consider the following two-dimensional model

$$\frac{\mathrm{d}x}{\mathrm{d}t} = ay, \qquad \frac{\mathrm{d}y}{\mathrm{d}t} = bx, \qquad (18)$$

which is a linear system with respect to both the parameter $\theta = (a, b)^{\mathrm{T}}$ and the state variable $(x, y)^{\mathrm{T}}$. Assume the

observational data of x and y and their time derivatives $\dot{x} := \mathrm{d}x/\mathrm{d}t$ and $\dot{y} := \mathrm{d}y/\mathrm{d}t$ are available at time t_i for $i = 1, \ldots, I$. Define a matrix $M_i = \begin{pmatrix} y_i & 0 \\ 0 & x_i \end{pmatrix}$, which takes values at time t_i. To estimate the parameters θ, linear regression can be easily applied to find the least-squares solution,

$$\theta = \left(\sum_i M_i^\mathrm{T} M_i \right)^{-1} \left(\sum_i M_i^\mathrm{T} z_i \right), \qquad (19)$$

where $z_i = (\dot{x}_i, \dot{y}_i)^\mathrm{T}$. Writing down the component-wise form of (19) yields

$$a = \left(\sum_i y_i^2 \right)^{-1} \left(\sum_i y_i \dot{x}_i \right). \qquad (20)$$

Now consider a different situation, where only x_i and $\dot{x}_i$ are observed. This is typical in many applications where only partial observations are available. Since estimating the parameter a requires the information of y as shown in (20), the state of y at each time t_i needs to be estimated, which naturally introduces uncertainty. Assume the estimated state of y_i is given by a Gaussian distribution, and y_i is written into the Reynolds decomposition form as $y_i = \langle y_i \rangle + y_i'$. The parameter estimation problem can be regarded as repeatedly drawing samples from the distribution of y_i, plugging them into (20), and then taking the average for the evaluation of the terms involving y_i to reach the maximum likelihood solution. In light of the fact from (13) that $\langle y_i^2 \rangle = \langle y_i \rangle^2 + \langle (y_i')^2 \rangle$, the above procedure essentially gives the following modified version of (20),

$$a = \left(\sum_i \left(\langle y_i \rangle^2 + \langle (y_i')^2 \rangle \right) \right)^{-1} \left(\sum_i \langle y_i \rangle \dot{x}_i \right), \qquad (21)$$

where the term $\langle (y_i')^2 \rangle$ arrives due to the average of the nonlinear function y_i^2 in (20). This is somewhat surprising at a glance since the underlying model (18) is linear. Nevertheless, the formula in (21) is intuitive. Consider estimating a from $\dot{x} = ay_i$ with data at only one time instant t_i. If $y_i \neq 0$ is deterministic, then $a = \dot{x}/y_i$. If y_i contains uncertainty and satisfies a Gaussian distribution, then the reciprocal distribution $\dot{x}/y_i$ is no longer Gaussian. The variance of y_i affects the mean value $\langle \dot{x}/y_i \rangle$, which is different from $\dot{x}/\langle y_i \rangle$. Following the same logic, as y_i^2 on the right-hand side of the least-squares solution (20) is nonlinear, the additional term $\langle (y_i')^2 \rangle$ appears in (20), which affects the parameter estimation skill. Therefore, uncertainty can play a significant role even in diagnosing a linear system where nonlinearity appears in the diagnostics (parameter estimation formula).

As a numerical illustration, assume enough data is generated from (18) with the true parameters $a = 2$ and $b = -2$. When the uncertainty, namely the variance of estimating y_i, is given by $\langle (y_i')^2 \rangle = 0.5, 1$ and 2 at all times t_i, the estimated parameter of a becomes $a = 1.000, 0.672$, and 0.404, respectively. As the estimated parameter becomes less accurate, the residual term in the regression increases, accounting for the effect of the input uncertainty.

In practice, expectation-maximization (EM) iterative algorithms can be applied to alternatively update the estimated parameter values and recover the unobserved states of y [Che23]. In the E step, the state of y can be inferred via DA, using the current parameter values θ. In the M step, the maximum likelihood estimate is applied to infer the parameters given the estimated state of y. Likewise, data augmentation can be used to sample the trajectories of the unobserved variables. This can be incorporated into other parameter estimation methods, such as the Markov Chain Monte Carlo [PRS13].

Eddy identification. Let us now explore how simple UQ tools advance the study of realistic problems. Oceanic eddies are dynamic rotating structures in the ocean. The primary goal is to show that nonlinearity in the eddy diagnostic will make UQ play a crucial role. Providing a rigorous definition of an eddy and discussing the pros and cons of different eddy identification methods are not the main focus here. Mesoscale eddies are major drivers of the transport of momentum, heat, and mass, as well as biochemical and biomass transport and production in the ocean. The study of ocean eddies is increasingly important due to climate change and the vital role eddies play in the rapidly changing polar regions and the global climate system.

Due to the complex spatial and dynamical structure of eddies there is no universal criterion for eddy identification. The Okubo-Weiss (OW) parameter [Oku70, Wei91] is a widely used approach based on physical properties of the ocean flow:

$$\mathrm{OW} = s_\mathrm{n}^2 + s_\mathrm{s}^2 - \omega^2, \qquad (22)$$

where the normal strain, the shear strain, and the relative vorticity are given by

$$s_\mathrm{n} = u_x - v_y, \quad s_\mathrm{s} = v_x + u_y, \quad \text{and} \quad \omega = v_x - u_y,$$

respectively, with $\mathbf{u} = (u, v)$ the two-dimensional velocity field and the shorthand notation $u_x := \partial u/\partial x$. When the OW parameter is negative, the relative vorticity is larger than the strain components, indicating vortical flow. The OW parameter is an Eulerian quantity based solely on a snapshot of the ocean velocity field. There are other quantities to identify eddies, such as the Lagrangian descriptor [VKGF16], based on a sequence of snapshots.

The use of these eddy identification diagnostics requires knowing the exact flow field. However, uncertainties may appear in state estimation, for example, in the marginal ice zone when the ocean is estimated via the LaDA using a
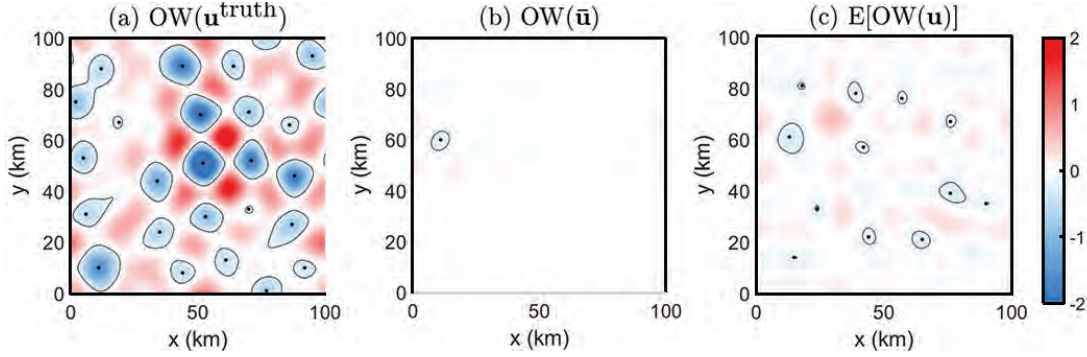
Figure 6. Eddy identification using the OW parameter. Panel (a): the OW parameter based on the true flow field. Panel (b): the OW parameter based on the estimated mean flow field from LaDA using a small number of the observed tracers. Panel (c): the expected value of the OW parameter based on multiple flow fields sampled from the posterior distribution.

limited number of ice floe trajectories. Eddy identification exploiting the mean estimate of the flow field can lead to large biases.

The contribution of the uncertainty in affecting the OW parameter can be seen by applying the Reynolds decomposition to each component in (22), e.g., $u_x = \bar{u}_x + u'_x$. By sampling multiple realizations from the posterior distribution of the estimated flow field, the expectation of the OW parameter values applying to each sampled flow field is given by

$$\mathbb{E}[\text{OW}(\mathbf{u})] = \text{OW}(\bar{\mathbf{u}}) + \langle (u'_x)^2 \rangle - 2\langle u'_x v'_y \rangle \\ + \langle (v'_y)^2 \rangle + 4\langle v'_x u'_y \rangle. \tag{23}$$

On the right-hand side of (23), $\text{OW}(\bar{\mathbf{u}})$ is the OW parameter applying to the estimated mean flow field. The remaining terms are all nonlinear functions of the fluctuation part due to the uncertainty in state estimation. In general, as most of the eddy identification criteria are nonlinear with respect to the flow field as in (22), changing the order of taking the expectation and applying the eddy diagnostic criterion will significantly impact the results. Notably, even though (21) and (23) are in entirely different contexts, they share the same essence that uncertainty plays a significant role when nonlinear terms appear in the diagnostics.

Panels (b)–(c) of Figure 6 numerically illustrate the difference between $\mathbb{E}[\text{OW}(\mathbf{u})]$ and $\text{OW}(\bar{\mathbf{u}})$, where the flow field is inferred from LaDA using a small number of the observed tracers. In the presence of such a large uncertainty, the main contribution of the estimated flow field comes from the fluctuation part, while the mean estimation has weak amplitudes. Consequently, very few eddies are identified using $\text{OW}(\bar{\mathbf{u}})$. The expected value $\mathbb{E}[\text{OW}(\mathbf{u})]$ improves the diagnostic results. However, it is still far from the truth in Panel (a) due to the large uncertainty of $\text{OW}(\mathbf{u})$ computed by using different realizations of $\mathbf{u}$ sampled from the posterior distribution of the LaDA. This is again similar to the intrinsic inaccuracy in the parameter estimation problem (21) when uncertainty appears. No-

tably, collecting $\text{OW}(\mathbf{u})$ associated with different sampled flow realizations allows us to compute the PDF $p(\text{OW}(\mathbf{u}))$, which provides a probabilistic view of eddy identification. Such a probabilistic eddy identification framework allows for assigning a probability of the occurrence of each eddy and the PDFs of lifetime and size of each eddy [CCWL24].

UQ in Advancing Efficient Modeling

The governing equations of many complex turbulent systems are given by nonlinear partial differential equations (PDEs). The spectral method remains one of the primary schemes for finding numerical solutions. For simplicity, consider a scalar field $u(x, t)$, which satisfies a PDE with quadratic nonlinearities, as in many fluids and geophysical problems. Assume Fourier basis functions $\exp(ikx)$ are utilized, where k is the wavenumber, and x is the spatial coordinate. Denote by $\hat{u}_k(t)$ the time series of the spectral mode associated with wavenumber k. Apply a finite-dimensional truncation to retain only the modes $k \in [-K, K]$. The resulting equation of $\hat{u}_k(t)$ typically has the following form,

$$\frac{d\hat{u}_k(t)}{dt} = \left(-d_k + i\omega_k \right) \hat{u}_k(t) + \hat{f}_k(t) \\ + \sum_{-K \leq m \leq K} c_{m,k} \hat{u}_m \hat{u}_{k-m}, \tag{24}$$

where $K \gg 1$ represents the resolution of the model, i is the imaginary unit, d_k and ω_k are both real, while $\hat{f}_k(t)$ is complex. The first term on the right-hand side of (24) is linear, representing the effect of damping/dissipation $-d_k < 0$ and phase ω_k. The second term is deterministic forcing. The last term sums over all the quadratic nonlinear interactions projected to mode k. Typically, a large number of such nonlinear terms will appear in the summation, which is one of the main computational costs in solving (24).

Due to the turbulence of these systems, many practical tasks, such as the statistical forecast and DA, require obtaining a forecast distribution by repeatedly running the gov-
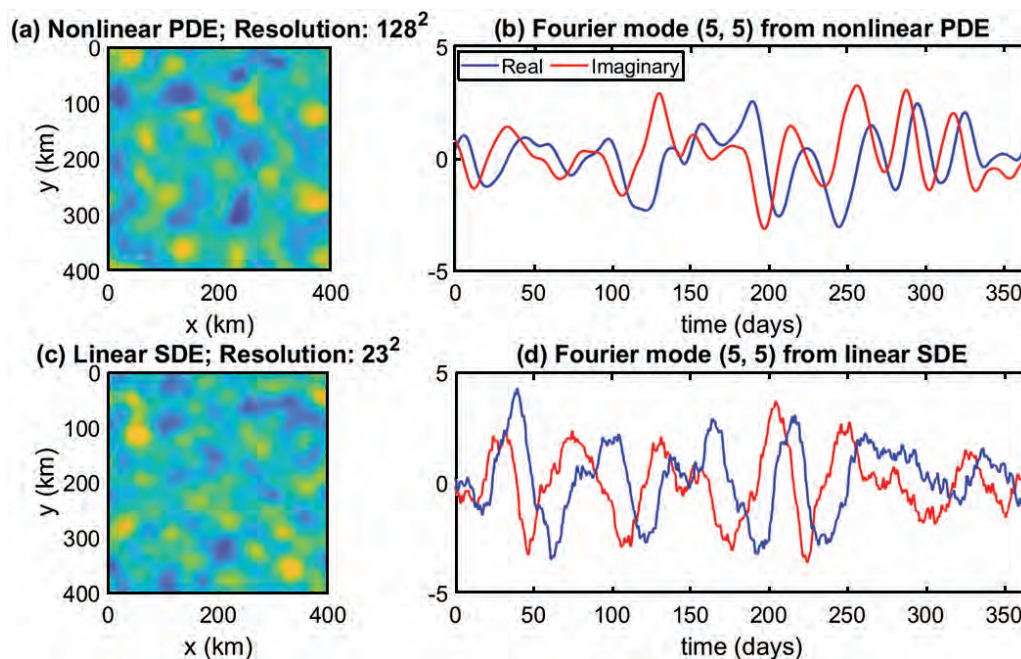
Figure 7. Comparison of the simulations from a nonlinear PDE model and a stochastic model, showing the latter can reproduce the forecast statistics of the former. Panels (a) and (c): snapshots of stream functions of the upper-layer ocean. Panels (b) and (d): time series of mode $(5, 5)$. These are two different realizations, so there is no pointwise correspondence between them in the snapshots or time series.

erning equation. Each run is already quite costly, so the ensemble forecast is usually computationally prohibitive. Therefore, developing an appropriate stochastic surrogate model is desirable, aiming to significantly accelerate computational efficiency. Stochasticity can mimic many of the features in turbulent systems. Since the goal is to reach the forecast statistics instead of a single individual forecast trajectory, once appropriate UQ is applied to guide the development of the stochastic surrogate model, it can reproduce the statistical features of the underlying nonlinear deterministic system.

In many applications, a large portion of the energy is explained by only a small number of the Fourier modes. Yet, the entire set of the Fourier modes has to be solved together in the direct numerical simulation to guarantee numerical stability. Therefore, reducing the computational cost by using stochastic surrogate models is twofold. First, in the governing equation of each mode, the heavy computational burden of calculating the summation of a large number of nonlinear terms will be replaced by computing a few much cheaper stochastic terms. Second, as the governing equation of each mode becomes independent when the nonlinear coupling is replaced by the stochastic terms, only the leading few Fourier modes need to be retained in such an approximate stochastic system, saving a large amount of computational storage. This also allows a larger numerical integration time step since stiffness usually comes from the governing equations of the small-scale

modes. One simple stochastic model is to replace all the nonlinear terms in the governing equation of each mode by a single stochastic term $\dot{W}$ representing white noise, which works well if the statistics of the mode are nearly Gaussian. The resulting stochastic model reads [Gar04]:

$$\frac{d\hat{u}_k(t)}{dt} = \left(-d_k + i\omega_k \right)\hat{u}_k(t) + \hat{f}_k(t) + \sigma_k \dot{W}_k. \quad (25)$$

For systems with strong non-Gaussian statistics, other systematic methods can be applied to develop stochastic surrogate models [Che23].

UQ plays a crucial role in calibrating the stochastic surrogate model. Specifically, the parameters in the stochastic surrogate model are optimized so that the two models have the same forecast uncertainty, which is crucial for DA and ensemble prediction. This is often achieved by matching a set of key statistics of the two models, especially the equilibrium PDFs and the decorrelation time. For (25), simple closed analytic formulae are available for model calibration in reproducing the forecast uncertainty, allowing it to be widely used in many practical problems. See the online supplementary document for more details.

Panels (a) of Figure 7 shows a snapshot of the stream function from a two-layer quasi-geostrophic (QG) model, which is given by a set of nonlinear PDEs with the spatial resolution of each layer being 128×128. In comparison, Panel (c) shows a snapshot of the spatial field by running a set of stochastic surrogate models, each calibrated by capturing the forecast uncertainty of one Fourier mode

of the two-layer QG model [CCW22]. Although the spatial resolution is only 23×23, which contains less than 2% of the total number of modes, the spatial field is qualitatively similar to the QG model. Panels (b) and (d) show the time series of Fourier mode $(5, 5)$ associated with the two models. They behave similarly as they have the same temporal correlation and PDFs. An additional example of using the stochastic model (25) to approximate the KdV equation can be found in [MMQ19].

Conclusions

This paper exploits simple examples to provide the basic concepts of UQ. It does not aim to provide a single, comprehensive definition, as various communities currently have different interpretations of UQ. Nevertheless, the information quantities (1) and (6) are natural measurements to quantify the uncertainty. As a developing field within applied mathematics and interdisciplinary research, UQ continues to evolve. One of the goals of this paper is to offer new insights into how these ideas can be applied across different fields, helping to reveal the commonalities and practical advantages of diverse approaches. As uncertainty is ubiquitous, incorporating UQ into analysis or strategic planning is essential to facilitate understanding nature and problem-solving in almost all disciplines. UQ has become an important tool in tackling observational data [BHTG21], forecasting turbulent systems [MB12], and in studying inverse problems [DS11]. UQ is not a standalone subject. It strongly relates to and advances the development of new techniques for mathematical reasoning, computational algorithms, effective modeling, data analysis, machine learning, and various applications. The areas in science, engineering, and technology requiring careful UQ are rapidly expanding and demands new ideas for coping with challenging issues.

Supplementary document and codes availability. The arXiv version of this article includes a supplementary document (`https://arxiv.org/abs/2408.01823`), which contains a comprehensive tutorial of many numerical examples shown in this article and beyond. The codes for these examples, written in both MATLAB and Python, are available from GitHub at `https://github.com/marandmath/UQ_tutorial_code`.

References

[ABN16] Mark Asch, Marc Bocquet, and Maëlle Nodet, *Data assimilation*, Fundamentals of Algorithms, vol. 11, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2016. Methods, algorithms, and applications, DOI 10.1137/1.9781611974546.pt1. MR3602006

[AJS08] Amit Apte, Christopher KRT Jones, and AM Stuart, *A Bayesian approach to Lagrangian data assimilation*, Tellus A: Dynamic Meteorology and Oceanography **60** (2008), no. 2, 336–347.

[BHTG21] Amy Braverman, Jonathan Hobbs, Joaquim Teixeira, and Michael Gunson, *Post hoc uncertainty quantification for remote sensing observing systems*, SIAM/ASA J. Uncertain. Quantif. **9** (2021), no. 3, 1064–1093, DOI 10.1137/19M1304283. MR4296767

[CCW22] Jeffrey Covington, Nan Chen, and Monica M. Wilhelmus, *Bridging gaps in the climate observation network: A physics-based nonlinear dynamical interpolation of lagrangian ice floe measurements via data-driven stochastic models*, Journal of Advances in Modeling Earth Systems **14** (2022), no. 9, e2022MS003218.

[CCWL24] Jeffrey Covington, Nan Chen, Stephen Wiggins, and Evelyn Lunasin, *Probabilistic eddy identification with uncertainty quantification*, arXiv:2405.12342 (2024).

[Che23] Nan Chen, *Stochastic methods for modeling and predicting complex dynamical systems—uncertainty quantification, state estimation, and reduced-order models*, Synthesis Lectures on Mathematics and Statistics, Springer, Cham, 2023, DOI 10.1007/978-3-031-22249-8. MR4572946

[CMT14] Nan Chen, Andrew J. Majda, and Xin T. Tong, *Information barriers for noisy Lagrangian tracers in filtering random incompressible flows*, Nonlinearity **27** (2014), no. 9, 2133–2163, DOI 10.1088/0951-7715/27/9/2133. MR3247074

[CT06] Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, 2nd ed., Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006. MR2239987

[DS11] M. Dashti and A. M. Stuart, *Uncertainty quantification and weak approximation of an elliptic inverse problem*, SIAM J. Numer. Anal. **49** (2011), no. 6, 2524–2542, DOI 10.1137/100814664. MR2873245

[Gar04] C. W. Gardiner, *Handbook of stochastic methods for physics, chemistry and the natural sciences*, 3rd ed., Springer Series in Synergetics, vol. 13, Springer-Verlag, Berlin, 2004, DOI 10.1007/978-3-662-05389-8. MR2053476
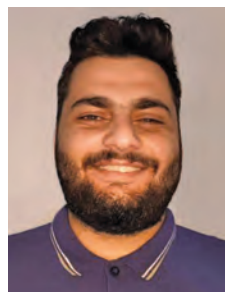
[Gra11] Robert M. Gray, *Entropy and information theory*, Springer Science & Business Media, 2011.

[Jef73] Harold Jeffreys, *Scientific inference*, Cambridge University Press, 1973.

[Kal60] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Trans. ASME Ser. D. J. Basic Engrg. **82** (1960), no. 1, 35–45. MR3931993

[Lor63] Edward N. Lorenz, *Deterministic nonperiodic flow*, J. Atmospheric Sci. **20** (1963), no. 2, 130–141, DOI 10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2. MR4021434

[LSZ15] Kody Law, Andrew Stuart, and Konstantinos Zygalakis, *Data assimilation*, Texts in Applied Mathematics, vol. 62, Springer, Cham, 2015. A mathematical introduction, DOI 10.1007/978-3-319-20325-6. MR3363508

[MB12] Andrew J. Majda and Michal Branicki, *Lessons in uncertainty quantification for turbulent dynamical systems*, Discrete and Continuous Dynamical Systems **32** (2012), no. 9, 3133–3221, https://www.aimsciences.org/article/id/4bc1cfda-d8e9-4f8c-b9d7-d82194279792.

[MKC02] Andrew Majda, Richard Kleeman, and David Cai, *A mathematical framework for quantifying predictability through relative entropy*, Methods Appl. Anal. **9** (2002), no. 3, 425–444, DOI 10.4310/MAA.2002.v9.n3.a8. Special issue dedicated to Daniel W. Stroock and Srinivasa S. R. Varadhan on the occasion of their 60th birthday. MR2023134

[MMQ19] Andrew J. Majda, M. N. J. Moore, and Di Qi, *Statistical dynamical model to predict extreme events and anomalous features in shallow water waves with abrupt depth change*, Proc. Natl. Acad. Sci. USA **116** (2019), no. 10, 3982–3987, DOI 10.1073/pnas.1820467116. MR3923854

[MQ18] Andrew J. Majda and Di Qi, *Strategies for reduced-order models for predicting the statistical responses and uncertainty quantification in complex turbulent dynamical systems*, SIAM Rev. **60** (2018), no. 3, 491–549, DOI 10.1137/16M1104664. MR3841156

[Mül06] Peter Müller, *The equations of oceanic motions*, Cambridge University Press, 2006.

[Oku70] Akira Okubo, *Horizontal dispersion of floatable particles in the vicinity of velocity singularities such as convergences*, Deep sea research and oceanographic abstracts, 1970, pp. 445–454.

[PRS13] Omiros Papaspiliopoulos, Gareth O. Roberts, and Osnat Stramer, *Data augmentation for diffusions*, J. Comput. Graph. Statist. **22** (2013), no. 3, 665–688, DOI 10.1080/10618600.2013.783484. MR3173736

[Sha48] C. E. Shannon, *A mathematical theory of communication*, Bell System Tech. J. **27** (1948), 379–423, 623–656, DOI 10.1002/j.1538-7305.1948.tb01338.x. MR26286

[VKGF16] Rahel Vortmeyer-Kley, Ulf Gräwe, and Ulrike Feudel, *Detecting and tracking eddies in oceanic flow fields: a Lagrangian descriptor based on the modulus of vorticity*, Nonlinear Processes in Geophysics **23** (2016), no. 4, 159–173.

[Wei91] John Weiss, *The dynamics of enstrophy transfer in two-dimensional hydrodynamics*, Phys. D **48** (1991), no. 2-3, 273–294, DOI 10.1016/0167-2789(91)90088-Q. MR1102165
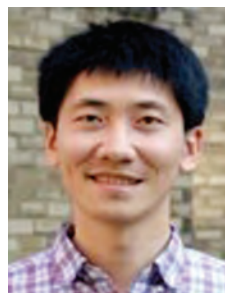
Nan Chen

Stephen Wiggins

Marios Andreou