Can Machine Learning Predict Extreme Events in Complex Systems?

Di Qi[a,1,2] **and Andrew J. Majda**[a,1,2]

[a]Department of Mathematics and Center for Atmosphere and Ocean Science, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012

This manuscript was compiled on October 3, 2019

Extreme events and the related anomalous statistics are ubiquitously observed in many natural systems, and it remains a grand challenge for the development of efficient methods to understand and accurately predict such representative features. Here, we investigate the skill of deep learning strategies in the prediction of extreme events in complex turbulent dynamical systems. Deep neural networks have been successfully applied to many imaging processing problems involving big data, and have recently shown potential for the study of dynamical systems. We propose to use a densely connected mixed-scale network model to capture the extreme events appearing in a truncated KdV (tKdV) statistical framework which creates anomalous skewed distributions consistent with recent laboratory experiments for shallow water waves across an abrupt depth change, where a remarkable statistical phase transition is generated by varying the inverse temperature parameter in the corresponding Gibbs invariant measures. The neural network is trained using data without knowing the explicit model dynamics, and the training data is only drawn from the near-Gaussian regime of the tKdV model solutions without the occurrence of large extreme values. A relative entropy loss function together with empirical partition functions is proposed for measuring the accuracy of the network output where the dominant structures in the turbulent field are emphasized. The optimized network is shown to gain uniformly high skill in accurately predicting the solutions in a wide variety of statistical regimes including highly skewed extreme events. The new technique is promising to be further applied to other complicated high-dimensional systems.

anomalous extreme events | convolutional neural networks | turbulent dynamical systems

Extreme events and their anomalous statistics are ubiquitous in various complex turbulent systems such as the climate, material and neuroscience, as well as engineering design (1–4). Understanding and accurate prediction of such phenomena remain a grand challenge, and have become an active contemporary topic in applied mathematics (5–8). Extreme events can be isolated rare events (2, 9, 10), or they can often be intermittent and even frequent in space and time (6, 8, 11, 12). The curse of dimension forms one important obstacle for the accurate prediction of extreme events in large complex systems (3, 4, 6, 13), where both novel models and efficient numerical algorithms are required. A typical example can be found in recent laboratory experiments for turbulent surface water waves going through an abrupt depth change revealing a remarkable transition to anomalous extreme events from near-Gaussian incoming flows (1).

A statistical dynamical model is then proposed in (14, 15) that successfully predicts the anomalous extreme behaviors observed in the shallow water wave experiments. The truncated Korteweg-de Vries (tKdV) equation is proposed as the governing equation modeling the flow surface displacement. Gibbs invariant measures are induced based on the Hamiltonian form of the tKdV equation to describe the probability distributions at equilibrium. A statistical transition from symmetric near-Gaussian statistics to a highly skewed probability density function (PDF) is achieved by simply controlling the 'inverse temperature' parameter in the Gibbs measure (15).

In recent years, machine learning strategies, particularly the deep neural networks, have been extensively applied to a wide variety of problems involving big data, such as image classification and identification (16–19). On the other hand, it still remains an actively growing topic to construct proper deep learning strategies for the study of complex turbulent flows. The deep neural network tools developed for imaging processing have been suggested to be applied for data-driven predictions of chaotic dynamical systems (20, 21), climate and weather forecasts (22, 23), and parameterization of unresolved processes (24–26). In the statistical prediction of extreme events, the available data for training is often restricted in limited regimes. A successful neural network is required to maintain adaptive skill in wider statistical regimes with vastly distinct statistics away from the training dataset. Besides, a working prediction model for turbulent systems would also require the prediction time scale longer than the decorrelation time that characterizes the mixing rate of the state variables.

In this paper, we investigate the extent of skill of the deep neural networks in predicting statistical solutions of complex turbulent systems, especially involving highly skewed probability density functions. The statistical tKdV equations serve as a difficult first test model for extreme event prediction with

Significance Statement

Understanding and predicting extreme events as well as the related anomalous statistics is a grand challenge in complex natural systems. Deep convolutional neural networks provide a useful tool to learn the essential model dynamics directly from data. A new deep learning strategy is proposed to predict the extreme events appeared in turbulent dynamical systems. A truncated KdV model displaying distinctive statistics from near-Gaussian to highly skewed distributions is used as the test model. The neural network is trained using data only from the near-Gaussian regime without the occurrence of large extreme values. The optimized network demonstrates uniformly high skill in successfully capturing the solution structures in a wide variety of statistical regimes including the highly skewed extreme events.

D. Q. and A. J. M. designed the research, performed the research, and wrote the paper.

The authors declare no conflict of interest.

[1]D. Q. and A. J. M. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: qidi@cims.nyu.edu, jonjon@cims.nyu.edu.

simple trackable dynamics but a rich variety of statistical regimes from near Gaussian to highly skewed PDFs showing extreme events. The important questions to ask are whether the deep networks can be trained to learn the complex hidden structures in the highly nonlinear dynamics purely from restricted data, and what are the essential structures required in the network to gain the ability to capture extreme events.

Our major goal here is to get accurate statistical prediction for the extreme events in time intervals significantly longer than the decorrelation time of the complex turbulent system. To achieve this, a convolutional neural network (MS-D Net) which exploits multi-scale connections and densely connected structures (27) is adopted to provide the basic network architecture to be trained using the model data from the tKdV equation solutions. This network enjoys the benefits of simpler model implementation and a smaller number of tunable training hyperparameters. Thus it becomes much easier to train requiring less computational cost and technical tuning of the hyperparameters.

The key structures for the neural network to successfully capture extreme events include: i) the use of a relative entropy (*Kullback-Leibler* divergence) loss function to calibrate the closeness between the target and the network output as distribution functions, so that the crucial shape of the model solutions are captured instead of a pointwise fitting in the turbulent output field values; and ii) calibrating the output data under a combination of empirical partition functions emphasizing the large positive and negative values in the model prediction so that the main features in the solutions are further emphasized. This convolutional neural network model enjoys the following major advantages for the prediction of extreme events:

- The simple basic network architecture makes the model easier to train and efficient to predict the solutions among different statistical scenarios, as well as reduces the danger of overfitting from data;
- The network structure approximates the original model dynamics with the designed model loss function and treatment of output data. So the complex system dynamics is easier to be learned from data;
- The temporal and spatial correlations in different scales are modeled automatically from the design of the network with convolution kernels representing different scales in different layers;
- The method shows robust performance with different model hyperparameters, and can be generalized for the prediction of more complicated turbulent systems.

Direct numerical tests show high skill of the neural network in successfully capturing the extreme values in the solutions with the model parameters only learned from the near-Gaussian regime of vastly different statistics. The model also displays accurate prediction in much longer time beyond the decorrelation time scale of the state, proving the robustness of the methods. The successful prediction in the tKdV equation implies the potential of future applications of the network to more complicated high dimensional systems.

Background for extreme events and neural network structures

The truncated KdV equations with extreme events. The tKdV model provides a desirable set of equations capable of capturing many complex features in surface water wave turbulence with simple trackable dynamics. Through a high wavenumber cutoff at Λ (with $J = 2\Lambda+1$ grid points), the Galerkin projected state $u = \sum_{1\leq|k|\leq\Lambda} \hat{u}_k(t)\, e^{ikx}$ induces stronger turbulent dynamics than the original continuous one (15). The tKdV equation has been adapted to describe the sudden phase transition in statistics (14) where highly skewed extreme events are generated from near-Gaussian statistics for waves propagating across an abrupt depth change. The tKdV model is formulated on a periodic domain $x \in [-\pi, \pi]$ as

$$u_t + E_0^{1/2} L_0^{-3/2} D_0^{-3/2} u u_x + L_0^{-3} D_0^{1/2} u_{xxx} = 0, \qquad [1]$$

where the state variable $u(x,t)$ represents the surface wave displacement to be learned directly using the deep neural network. The model is nondimensionalized using the characteristic scales E_0 as the total energy, L_0 as the length scale, and D_0 as the water depth. The steady state distribution of the tKdV solution can be described by the invariant Gibbs measure derived from the equilibrium statistical mechanics (28)

$$\begin{aligned}\mathcal{G}_\theta(u) = & C_\theta^{-1} \exp\Big\{-\theta \Big[E_0^{1/2} L_0^{-3/2} D_0^{-3/2} H_3(u) \\ & -L_0^{-3} D_0^{1/2} H_2(u)\Big]\Big\}\, \delta\left(E(u)-1\right),\end{aligned} \qquad [2]$$

with a competition between the cubic term $H_3 = \frac{1}{6}\int u^3$ and the quadratic term $H_2 = \frac{1}{2}\int u_x^2$ from the Hamiltonian. The last term in [2] is the delta function constraining the total energy conservation, $E(u) = \frac{1}{2}\int u^2 = 1$. The only parameter $\theta < 0$ as the 'inverse temperature' determines the skewness of the PDF of u (14). The Gibbs measures [2] with different values of θ can be used to provide initial samples for the direct simulations of the model [1]. Different final equilibrium statistics (with various skewness) can be obtained based on the initial configuration of the ensemble set (that is, from picking different inverse temperature values θ). A detailed description about the statistical tKdV model together with the simulation setup is provided in *SI Appendix, A*.

Training and prediction data from the same model with distinct statistics. The basic idea in training the deep neural network is to use a training dataset with solutions sampled from [2] using near-Gaussian statistics. The tKdV model dynamics can be learned from the training set without explicitly knowing the model dynamics. Then the question is what is the range of skill in the trained neural network to predict the highly skewed non-Gaussian distributions among different sets of data. The training and prediction datasets are proposed from the ensemble solutions of the tKdV model [1] based on the following strategy:

- In the training dataset, we generate solutions from an ensemble simulation starting from a near-Gaussian PDF using a small absolute value of the inverse temperature θ_0 (as shown in the first row and the near-Gaussian PDF in Figure 1). On one hand, the model dynamics is represented by the group of solutions $\{u_{\theta_0}\}$ to be learned

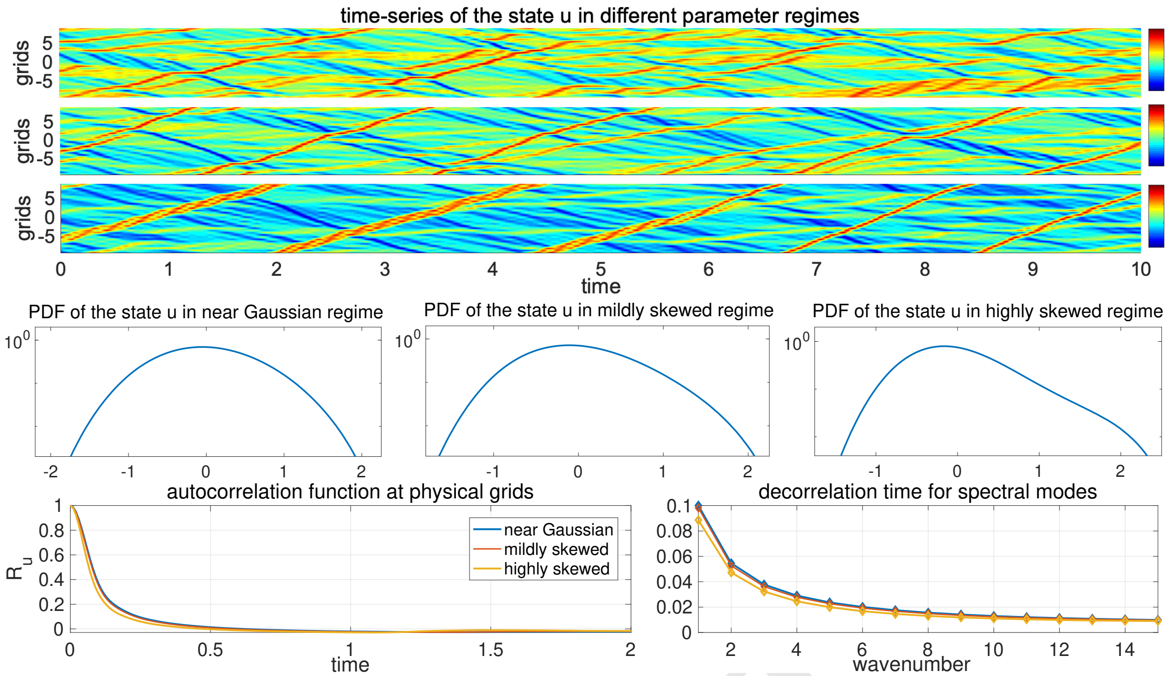


Fig. 1. Solutions and statistics of the tKdV equation in three typical parameter regimes with different statistics. The first three rows plot solution trajectories in the three regimes with near-Gaussian (first row), mildly skewed (second row), and highly skewed (third row) statistics. The corresponding equilibrium PDFs of the three cases are shown next. The autocorrelation functions and decorrelation time of each Fourier mode of the model state u are compared in the last row.

through the deep neural network. On the other hand, only near-Gaussian statistics is obtained in this training dataset so the neural network cannot know about the skewed rare events appearing in other statistical regimes directly from the training process.

- For the prediction dataset, we test the model skill using the data $\{u_\theta\}$ generated from various different initial inverse temperatures θ (as shown in the second and third rows as well as the skewed PDFs in Figure 1). It provides an interesting testbed to check the scope of skill in the optimized neural network for capturing the distinctive statistics and extreme events.

The choices of the training and prediction datasets are illustrated in Figure 1, which first shows in the first three rows realizations of the tKdV model solutions from different inverse temperatures θ. A smaller amplitude of θ gives near-Gaussian statistics in the model state u, while larger amplitudes of θ give trajectories with skewed PDFs. The corresponding equilibrium PDFs of the state u from direct ensemble dynamical solutions are compared next to illustrate explicitly the transition in statistics. Turbulent dynamics with multiscale structures are observed in all the tKdV solutions. The autocorrelation function $\mathcal{R}_u(t) = \langle u(t)\, u(0) \rangle$ and the decorrelation time $T_{\mathrm{decorr}} = \int \mathcal{R}_u(t)$ are plotted in the bottom row of Figure 1, confirming the rapid mixing in the solutions.

Data structures for the deep neural networks. The deep convolutional networks can be viewed as a function $\mathbf{y} = f_M(\mathbf{x})$, mapping the input signal $\mathbf{x} \in \mathbb{R}^{m\times n\times c}$ with m rows, n columns, and c channels to the output data $\mathbf{y} \in \mathbb{R}^{\mathrm{m}'\times \mathrm{n}'\times \mathrm{c}'}$ with m' rows, n' columns, and c' channels. We consider an ensemble simulation of M trajectories of the tKdV equation evaluated at the J grid points and measured in a time interval $[t_0, t_{N-1}]$. Thus the input data for the network comes from the ensemble solutions at the first N time measurements at $t_j = j\Delta t, j = 0, \cdots, N-1$,

$$\mathbf{x}^{(l)} = \left[u_0^{(l)}, u_1^{(l)}, \cdots, u_{J-1}^{(l)}\right]^{\mathrm{T}} (t_0, \cdots, t_{N-1}) \in \mathbb{R}^{J\times N},$$

as a tensor with $m = J$ rows for the spatial grid points, $n = N$ columns for the discretized time evaluations, and only one channel $c = 1$ for each of the input samples $l = 1, \cdots, M$. An ensemble of total M independent solutions from the Monte-Carlo simulation is divided into mini-batches to feed in the network in the training process. For simplicity, the output data is designed in the same shape as the predicted states evaluated at a later time $t = T + t_0$ starting from the previous initial data

$$\mathbf{y}^{(l)} = \left[u_0^{(l)}, u_1^{(l)}, \cdots, u_{J-1}^{(l)}\right]^{\mathrm{T}} (T + t_0, \cdots, T + t_{N-1}) \in \mathbb{R}^{J\times N}.$$

The forwarding time T controls how long we would like the network to push forward the states u in one time update. For one effective neural network for the complex system, the time scale T is expected to be longer than the decorrelation time $T > T_{\mathrm{decor}}$. The above construction is supposed to feed both the time and spatial correlations of the original dynamical model into the neural network to be learned in the approximation map $\mathbf{y} = f_M(\mathbf{x})$.

Deep convolutional neural network architecture. The basic structures of the convolutional neural network include the operators in each single convolution layer; and the connections between multiple layers. We would like to first keep the neural network in its simplest standard setup, so that we are able to concentrate on the improvement in key structures without risking at getting lost in manipulating the various complicated *ad hoc* hyperparameters. More detailed convolutional network construction is described in *SI Appendix, B.1* following the general neural network architecture as in (19, 27).

Basic convolutional neural network unit. In each single convolution layer, the input data from the previous layer output is updated in the general form

$$\mathbf{y}=\sigma\left(g_h\left(\mathbf{x}\right)+b\right).$$

A convolutional operator g_h is first applied on the input data $\mathbf{x}$ in a small symmetric window with size $w\times w$, where the first dimension controls the correlation in the spatial direction and the second one for the temporal correlation. A bias b is added to the convolved data before applying a final nonlinear operator σ using the common choice of rectified linear unit (ReLU) function. The convolution kernel starts with a small size 3×3 (that is, using only the two nearest neighbor points in space and time) which enables fast computation and easy control. Naturally, periodic boundary condition is applied on the spatial dimension and replicate boundary is added in time before t_0 and after t_{N-1} for the boundary padding. No additional structures are implemented in the convolution layer unit to keep the basic standard architecture used in imaging processing (19).

A densely connected and mix-scale structure. Next, we need to propose the connection between different layers. The common *feedforward deep neutral network* feeds the input data in the i-th layer only to the next $(i+1)$-th layer. The feedforward network often requires a larger number of layers to work, thus it is expansive to train and difficult to handle. Proper downscaling and upscaling steps going through the layers may also be required, while these downscaling and upscaling operations may not be a feasible approach for simulating the dynamical model time integration steps.

As an alternative approach, a *mixed-scale dense neural network* (MS-D Net) is introduced in (27) by mixing different scales within each layer using a dilated convolution, and densely connecting all the feature maps among all the layers. First, to learn the multiscale structures, the convolution kernels in different layers are dilated differently by adding s zeros between the values in the original kernel $w\times w$. The dilated convolutions become especially appealing for capturing the multiscale structures in the turbulent dynamics. Different spatial and temporal scales are included adaptively with different convolution length scales. Second, the dense network connection includes all the previous layer information to update the output data in the next layer. In the implementation, all the previous layer outputs are piled together as input channels for the next layer. Together with the multiscale convolution kernels used in different layers, the output in the next layer combines the information in different scales and produces a balanced update in the next step.

The mixed-scale dense neural network requires fewer feature maps and trainable parameters, so it is easier to handle compared with the direct feedforward network. It provides a desirable setup for the prediction in dynamical systems by feeding in all the data in previous layers decomposed into different scale structures. Then information at different scales communicates with each other through the dense network connection. Intuitively, this is a reasonable structure for the turbulent solutions since all the history information is useful for the prediction in the next steps. The densely connected network structure is also comparable to the time integration scheme, where all the history information is used to update the state at the next time step without using any downscaling and upscaling steps.

A new learning strategy for extreme event prediction

In this section, we construct the specific network structures designed for learning turbulent system dynamics then the prediction of extreme events. In this case with data from the turbulent models, small fluctuations in the solutions may introduce large errors in optimizing the loss function. We aim at capturing the dominant emerging features such as extreme events and are more interested in the statistical prediction rather than the exact locations of the extreme values.

Loss function calibrating the main dynamical features. In the training process, the model parameters are achieved through the optimization for the proper loss function (or cost) proposed depending on the target to be captured using the network. The two popular standard choices for the loss functions are using the L_1 error and the mean square L_2 error:

- L_1 *error loss*: this criterion measures the mean absolute error between each element in the output data $\mathbf{x}$ and target $\mathbf{y}$ through the L_1 distance

$$l_1\left(\mathbf{x},\mathbf{y}\right)=\frac{1}{M}\sum_{m=1}^{M}\left\|\mathbf{x}^{(m)}-\mathbf{y}^{(m)}\right\|, \qquad [3]$$

  where M is the training data size in one cycle, and $\left(\mathbf{x}^{(m)},\mathbf{y}^{(m)}\right)$ is one member of output data and target data from the training mini-batch.

- L_2 *error loss*: this criterion measures the mean squared error between each element in the output $\mathbf{x}$ and target $\mathbf{y}$ through the mean square L_2 norm

$$l_2\left(\mathbf{x},\mathbf{y}\right)=\frac{1}{M}\sum_{m=1}^{M}\left\|\mathbf{x}^{(m)}-\mathbf{y}^{(m)}\right\|^2, \qquad [4]$$

  among all the numbers of training samples M.

The above two loss functions offer pointwise measurements of the errors in space-time for each predicted sample from the network. This may cause problems especially when the system for prediction is highly turbulent with internal instability and is fast mixing. Small error perturbation in the input data may lead to vastly different solutions shortly after the mixing decorrelation time. The pointwise measurements focus on the accuracy at each value of the solutions, thus the small fluctuation errors might be accumulated and amplified under such metrics and unnecessary large weights are added to correct errors in the moderate-amplitude fluctuation parts.

On the other hand, we are most interested in the prediction of statistical features in the extreme events rather than the exact trajectory solutions of the system. The small shifts in the extreme value locations should be tolerated in the loss function. Therefore, a more useful choice could be the *relative entropy* loss function that measures the *Kullback-Leibler* divergence in the predicted density functions:

- *Relative entropy loss*: the relative entropy loss function computes the distance between two distribution functions

$$l_{\mathrm{KL}}\left(\tilde{\mathbf{x}},\tilde{\mathbf{y}}\right)=\frac{1}{M}\sum_{m=1}^{M}\sum_{i}\tilde{y}_{i}^{(m)}\log\frac{\tilde{y}_{i}^{(m)}}{\tilde{x}_{i}^{(m)}},\qquad[5]$$

  where the superscript m represents the mini-batch members to be measured in the relative entropy metric, and the subscript i goes through all the dimensions of the normalized variables $(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$ to be described next.

Under the relative entropy loss function in [5], the input data $\mathbf{x}$ and $\mathbf{y}$ are treated as distribution functions. The 'shapes' between the output data and the target are compared rather than the pointwise details, so that it guides the network to focus on the main model dynamical features instead of the turbulent fluctuations that are impossible to be fitted accurately. The additional difficulty to train the network using the relative entropy loss function is the constraint on the form of the output data to measure. The input of the relative entropy requires to be in the form of a density distribution function.

Scaling the output data with empirical partition functions. In this case of using the relative entropy loss function, we need to propose proper preprocessing of the output and target data to fit the required structure as a probability distribution. One direct way to do so is by taking the *softmax* function for the the output data from the neural network normalized by a partition function

$$\tilde{x}_{i}=\frac{\exp\left(x_{i}\right)}{\sum_{i}\exp\left(x_{i}\right)},\qquad[6]$$

before measuring the error in the loss function. In this way, the data to put into the relative entropy loss function is normalized inside the range $[0,1]$ with summation 1. This agrees with the definition in the relative entropy inputs. More importantly, this normalization emphasizes the large positive values of the data. Thus it offers a better calibration for the occurrence of positive major flow structures to be captured in the solutions.

Furthermore, a better choice for balancing both the positive and negative dominant values in the training data is to introduce scales with 'temperatures'. We use the following two empirical partition functions with both positive and negative coefficients to rescale the output data as

$$\tilde{x}_{i}^{+}=\frac{\exp\left(x_{i}/T_{+}\right)}{\sum_{i}\exp\left(x_{i}/T_{+}\right)},\quad\tilde{x}_{i}^{-}=\frac{\exp\left(-x_{i}/T_{-}\right)}{\sum_{i}\exp\left(-x_{i}/T_{-}\right)},\qquad[7]$$

where $T_{+}>0,T_{-}>0$ are the positive and negative temperatures weighing the importance of dominant large amplitude features in the scaled measure. Accordingly, the loss function to minimize under the relative entropy metric becomes a combination with the two empirical partition functions

$$l_{\mathrm{EPF}}\left(\mathbf{x},\mathbf{y}\right)=l_{\mathrm{KL}}\left(\tilde{\mathbf{x}}^{+},\tilde{\mathbf{y}}^{+}\right)+\alpha l_{\mathrm{KL}}\left(\tilde{\mathbf{x}}^{-},\tilde{\mathbf{y}}^{-}\right),\qquad[8]$$

where we use $\alpha>0$ as a further balance between the positive and negative temperature components. In this combined empirical partition function metric using [7] and [8], the major flow structures in the turbulent field represented by the dominant extreme values are better characterized from both the positive and negative sides in the statistics of the model. In the following computational experiments, we always pick $T_{+}=T_{-}=1$ and $\alpha=1$ for simplicity. More discussion for role of the balance weight α is shown in *SI Appendix, B.2*.

Model performance in training and prediction stages

Using the previous model construction, we show the training and prediction performance using the MS-D Net combined with the relative entropy loss function and rescaled output data using empirical partition functions applied on the tKdV equations. In the numerical tests, we first consider the optimal prediction skill using the deep neural network within one updating cycle. Especially, we are more interested in capturing the non-Gaussian statistics from the network rather than the exact recovery of the single time-series which should give large difference with small perturbations due to its turbulent nature. The basic strategy is to train the model using data from the near-Gaussian solutions with a small inverse temperature θ_0 in the Gibbs measure [2]; then the optimized neural network is used to predict the more skewed model statistics for regimes with larger absolute values of θ.

For simplicity, we set the input and output data of the network in the same shape. Then the model output from one single iteration of the network is compared with the target data from the true model solution. In the structure of the neural network, the standard setup is adopted for the tKdV solutions. In each layer of the network, a kernel with minimum size $w=3$ is taken for the convolution update. The dilation size for mixed scales changes from $s=0$ to $s=5$ repeatedly as the network grows in depth. Different network depths of layers L are tested for the model performance, while it is found that a moderate choice $L=80$ is enough to produce desirable training and prediction results.

In calibrating the errors from model predictions, we propose the normalized square error between the true target $\mathbf{y}^{\mathrm{t}}$ and the network output $\mathbf{y}^{\mathrm{o}}$

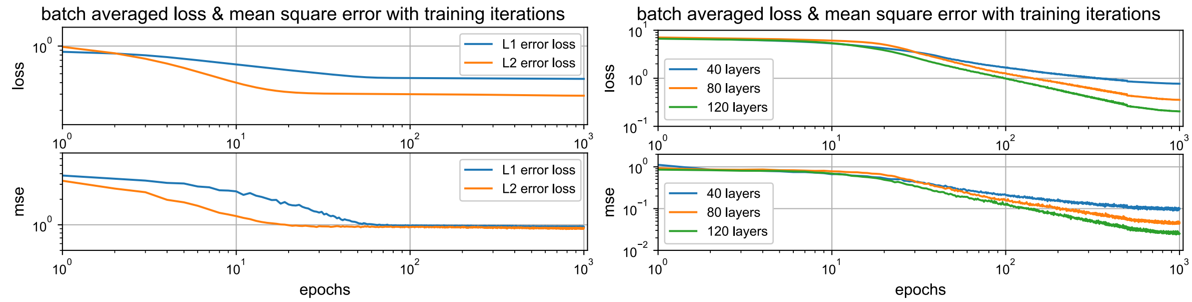
$$E\left(\mathbf{y}^{\mathrm{t}},\mathbf{y}^{\mathrm{o}}\right)=\frac{\sum_{i=1}^{JN}\left|f\left(y_{i}^{\mathrm{t}}\right)-f\left(y_{i}^{\mathrm{o}}\right)\right|^{2}}{\sum_{i=1}^{JN}\left|f\left(y_{i}^{\mathrm{t}}\right)\right|^{2}},\qquad[9]$$

where the subscript i represents the i-th component in the training/prediction set $\mathbf{y}\in\mathbb{R}^{J\times N}$. The function $f\left(y\right)$ acting on each component of $\mathbf{y}$ can be used to extract the useful features to be calibrated. In the following tests, we use $f\left(y\right)=y$ to compare the original output of the data, and use the exponential scaling $f\left(y\right)=\exp\left(y\right)$ to check the prediction in positive extreme values.

Training the neural network using near-Gaussian data. In the training process for turbulent system statistics, we include the temporal and spatial correlations together in the input data by considering a short time-series of the solution. The training data is drawn from the model solutions of [1] only among the near-Gaussian regime statistics. In summary, we use the training data set in the following structure:

- The *input data* is from the ensemble solutions $u_{j}^{(m)}\left(t_{n}\right)$ of the tKdV equation. It is organized in a tensor of the shape (M,J,N), where $M=10000$ is the total ensemble size, $J=32$ is the spatial discretization size, and $N=50$ is the sampled time instants with the time step $\Delta t=0.01$. Thus the initial data for training is given as the tKdV solutions in the time window $[0,0.5]$.

- The *target data* for the training result is the solutions $u_{j}^{(m)}\left(T+t_{n}\right)$ of the tKdV solution. The data is organized into the tensor with the same size (M,J,N) as the input



(a) L_1 and L_2 error as the loss function
(b) relative entropy loss with different layers

Fig. 2. Training loss function and the mean relative square error in the data using L_1 and L_2 error loss functions (left) and using the relative entropy loss function with rescaled output data (right) during the training iterations. Both networks with L_1 and L_2 loss are set to have $L = 80$ densely connected layers; and the networks with the relative entropy loss are compared using $L = 40, 80, 120$ layers.

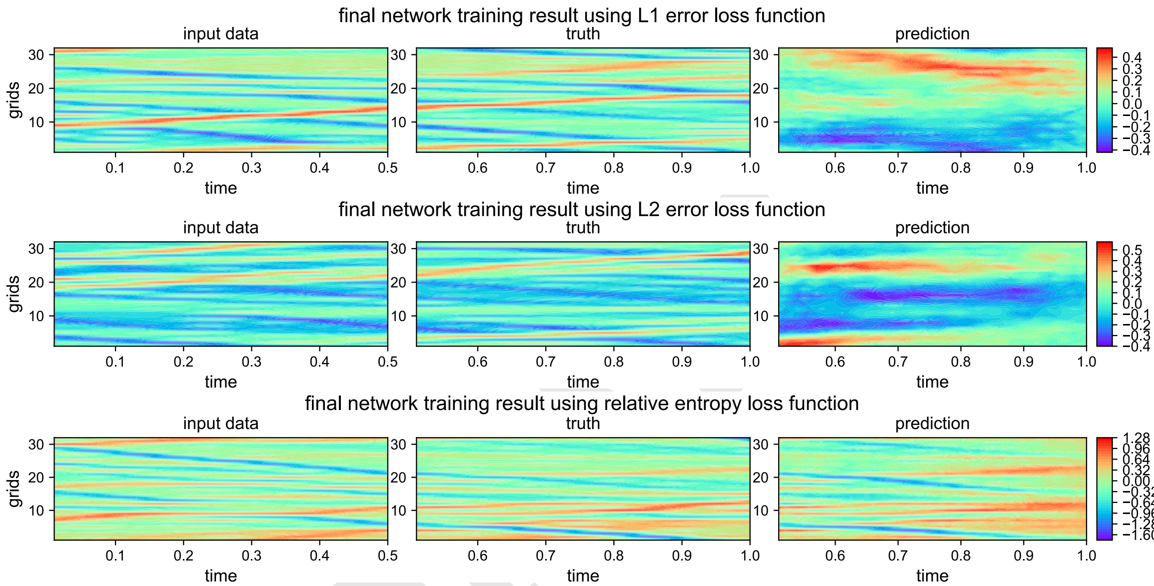


Fig. 3. One snapshot of the final training results with three different loss functions. The left panel shows the input data, the middle panel is the true target to fit, and the right panel shows the output results from the trained networks. All networks contain $L = 80$ layers in the tests.

data. We can consider different starting time for the target data by changing the time length T. The prediction time scale here is taken as $T = 0.5$, that is, to consider the prediction in time window $[0.5, 1]$.

- The input data (with total ensemble size 10000) is divided into 100 mini-batches with size 100 for each training group in one epoch. In total we use 1000 epochs in the entire training process. The mini-batches are randomly selected with the batch indices resampled from random numbers in each step.

Notice that the above prediction time length $T = 0.5$ used in the experiments is much longer beyond the decorrelation time of the tKdV states. From the direct numerical results in Figure 1, the autocorrelation function decays to near zero at $t = 0.5$, and the longest decorrelation time among the spectral modes is below 0.1.

First, we compare the training performance using the standard L_1 and L_2 loss functions in [3] and [4] with the same MS-D Net structure. The left panel of Figure 2 shows the evolution of training loss functions and the mean relative square errors [9] among the training samples during the stochastic gradient descent iterations. According to the loss functions under the L_1 and L_2 distances, the training appears to be effective and the error quickly dropped to smaller values in the first few steps. However, if we compare more carefully about the relative errors in the results, both cases get saturated quickly at a high error level near 1. The errors then become difficult to improve by training with larger number of iterations and applying deeper layers. This is because under both metrics, the model tries hard to fit the small-scale turbulent fluctuations in small values while missing the most important large-scale events in the solutions. It can be seen more clearly in Figure 3 for typical training output snapshots compared with the truth. No desirable prediction can be reached.

In contrast, significant improvement is achieved by switching to the relative entropy loss function and adopting empirical partition functions to normalize the output data. In this case as illustrated on the right panel of Figure 2, both the relative entropy loss function and the relative error drop to very small values in the final steps of the training iterations, implying high skill of the network to produce accurate predictions in the prediction time range (Though it appears that the accuracy

Table 1. Mean and variance of the relative square errors among a test with 500 samples for the state u and the scaled state $\exp(u)$.

error		near Gaussian	mildly skewed	highly skewed
u	mean	0.2682	0.2556	0.2690
	variance	0.0039	0.0048	0.0087
$\exp(u)$	mean	0.0733	0.0764	0.0985
	variance	0.0005	0.0011	0.0060

could be further improved by applying more iteration steps in the training, the results are already good enough after about 200 iterations). Note that we use logarithmic coordinates so the small values are emphasized. According to the last row of Figure 3 for one typical training result snapshot, both the extreme values and the small amplitude structures are captured in the model.

As a final remark, we check the proper depth needed for accurate predictions in the neural network. The right panel of Figure 2 also compares the same network under the relative entropy loss but using different numbers of layers. A deeper network clearly can further improve the prediction skill and push the final optimized error to an even lower value, with the cost of a larger computational requirement. Still from the comparison, it shows that a moderate number of layers (such as $L = 80$) is sufficient to produce accurate results with relatively low cost. By pushing the network to deeper layers with $L = 120$, the improvement in error just becomes small and may not be necessary with the additional computational cost. The last row of Figure 3 shows already the quite accurate recovery of the solution field purely learned from data.

Predicting extreme events using deep neural network. In checking the prediction skill of the optimized network, we pick the neural network with $L = 80$ densely connected layers as the standard model to test its predictions among different statistical regimes. It has been shown in the training process with a high skill in recovering the original flow structures. Next, we should confirm that the neural network has really learned the dynamical structure of the original model, instead of merely overfitting the data.

Three statistical regimes ranging from near Gaussian, mildly skewed, and highly skewed PDFs as shown in Figure 1 are taken for testing the range of prediction skill in the neural network model. An ensemble of 500 new trajectories from the tKdV solutions in different statistical regimes is used to show the robustness of the method. The mean and variance of the relative square errors [9] among the samples for the state u and the errors under the exponential function $\exp(u)$ are list in Table 1 for different statistical regimes. Uniform high accuracy in the mean with tiny variance is achieved among the vastly different regimes with distinct statistical features.

Especially, we are interested in the case with highly non-Gaussian statistics representing the frequent occurrence of extreme events. In Figure 4, the network is used to predict the flow solutions in the regime with highly skewed statistics (results for the other two cases can be found in *SI Appendix, B.2*). The extreme values are represented by high peaks of a dominant wave moving along the field. This feature is not shown at all in the training data where only near Gaussian statistics is presented. As shown in the results, the trained network displays uniform skill among all the tested samples in capturing the exact dynamical solutions in the extreme event domain unknown from the training data. By looking at the errors in the scaled data using the exponential function, the error amplitude even becomes smaller, confirming the accurate characterization of large extreme values through the network. In the typical snapshot of one sample, both the extreme values in the transporting waves and non-extreme detailed turbulent fluctuating structures are captured by the model.

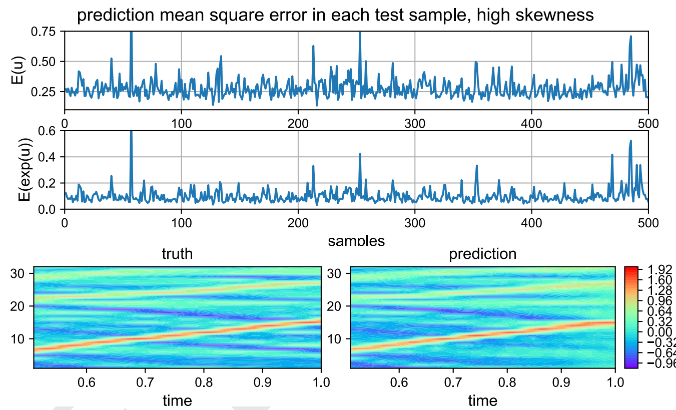


Fig. 4. Prediction in the regime with highly skewed statistics using the trained neural network with $L = 80$ layers. The upper row plots the relative square errors for the state u and the scaled state $\exp(u)$ among the 500 tested samples. The lower row shows one typical snapshot of the prediction.

Furthermore, the prediction errors as the absolute difference between the truth and the model prediction in the three statistical regimes are displayed in Figure 5. The network predictions maintain accurate with small errors in a much longer time scale than the decorrelation time $T_{\mathrm{decorr}} < 0.1$ of the states. This gives the final confirmation of the general high skill in the deep neural network for capturing key representative features in complex dynamical systems once the essential structures are learned from the training procedure.

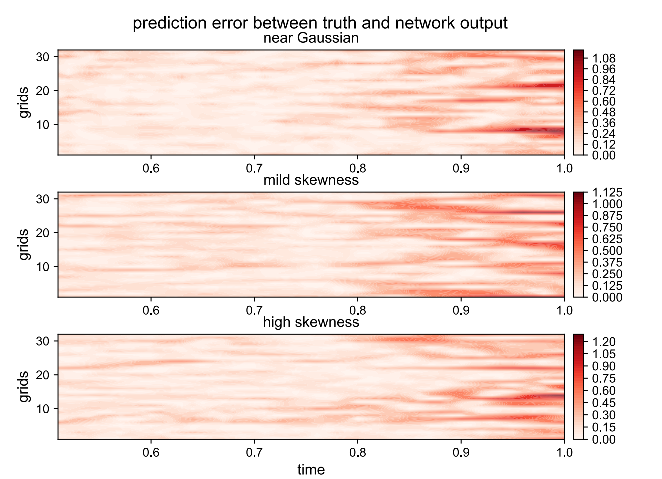


Fig. 5. The prediction error in the absolute difference between the truth and model output in the three tested regimes with different statistics.

Concluding remarks

A new strategy using a densely connected multi-scale deep neural network with relative entropy loss function for calibrating rescaled output data is proposed for the prediction of extreme events and anomalous features from data. It needs to be noticed that the extreme events are often represented by highly skewed PDFs and have frequent occurrence in the turbulent field (3, 8, 12) in contrast to the other situation of isolated rare events which can be studied with machine learning models (9). The prediction skill of the optimized deep neural network is tested on the truncated KdV equation, where different Gibbs states create a wide range of statistics from near Gaussian to highly skewed distributions. By adopting the densely connected and multi-scale structures, the deep neural network is easy to train with standard model setup and fewer model hyperparameters.

Using training data only drawn from the near-Gaussian regime of the dynamical model, the deep neural network displays high skill in learning the essential dynamical structures of the complex system and provides uniformly accurate prediction among a wide range of different regimes with distinct statistics. The network also shows robustness among tests in a large ensemble of samples. The robust performance in the test model implies the potential of more general applications using the neural network framework for the prediction of extreme events and important statistical features in a wider group of more realistic high-dimensional turbulent dynamical systems.

ACKNOWLEDGMENTS. This research of A.J.M. is partially supported by the Office of Naval Research N00014-19-S-B001. D.Q. is supported as a postdoctoral fellow on the grant.

1. Bolles CT, Speer K, Moore MNJ (2019) Anomalous wave statistics induced by abrupt depth change. *Physical Review Fluids* 4(1):011801.
2. Dematteis G, Grafke T, Vanden-Eijnden E (2018) Rogue waves and large deviations in deep sea. *Proceedings of the National Academy of Sciences* 115(5):855–860.
3. Qi D, Majda AJ (2018) Predicting extreme events for passive scalar turbulence in two-layer baroclinic flows through reduced-order stochastic models. *Communications in Mathematical Sciences* 16:17–51.
4. Mohamad MA, Sapsis TP (2018) Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* 115(44):11138–11143.
5. Qi D, Majda AJ (2016) Predicting fat-tailed intermittent probability distributions in passive scalar turbulence with imperfect models through empirical information theory. *Communications in Mathematical Sciences* 14(6):1687–1722.
6. Majda AJ, Tong XT (2015) Intermittency in turbulent diffusion models with a mean gradient. *Nonlinearity* 28(11):4171.
7. Viotti C, Dias F (2014) Extreme waves induced by strong depth transitions: Fully nonlinear results. *Physics of Fluids* 26(5):051705.
8. Cai D, Majda AJ, McLaughlin DW, Tabak EG (2001) Dispersive wave turbulence in one dimension. *Physica D: Nonlinear Phenomena* 152:551–572.
9. Guth S, Sapsis TP (2019) Machine learning predictors of extreme events occurring in complex dynamical systems. *Entropy* 21(10).
10. Cousins W, Sapsis TP (2015) Unsteady evolution of localized unidirectional deep-water wave groups. *Physical Review E* 91(6):063204.
11. Majda AJ, Lee Y (2014) Conceptual dynamical models for turbulence. *Proceedings of the National Academy of Sciences* 111(18):6548–6553.
12. Grooms I, Majda AJ (2014) Stochastic superparameterization in a one-dimensional model for wave turbulence. *Communications in Mathematical Sciences* 12(3):509–525.
13. Chen N, Majda AJ (2017) Beating the curse of dimension with accurate statistics for the Fokker–Planck equation in complex turbulent systems. *Proceedings of the National Academy of Sciences* 114(49):12864–12869.
14. Majda AJ, Moore MNJ, Qi D (2019) Statistical dynamical model to predict extreme events and anomalous features in shallow water waves with abrupt depth change. *Proceedings of the National Academy of Sciences* 116(10):3982–3987.
15. Majda AJ, Qi D (2019) Statistical phase transitions and extreme events in shallow water waves with an abrupt depth change. *submitted to Journal of Statistical Physics*.
16. Schmidhuber J (2015) Deep learning in neural networks: An overview. *Neural networks* 61:85–117.
17. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436.
18. Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. *Science* 349(6245):255–260.
19. Goodfellow I, Bengio Y, Courville A (2016) *Deep learning.* (MIT press).
20. Pathak J, Hunt B, Girvan M, Lu Z, Ott E (2018) Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Physical review letters* 120(2):024102.
21. Raissi M, Karniadakis GE (2018) Hidden physics models: Machine learning of nonlinear partial differential equations. *Journal of Computational Physics* 357:125–141.
22. Bolton T, Zanna L (2019) Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems* 11(1):376–399.
23. Weyn JA, Durran DR, Caruana R (2019) Can machines learn to predict weather? using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems.*
24. Han J, Jentzen A, E W (2018) Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences* 115(34):8505–8510.
25. E W, Han J, Jentzen A (2017) Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in Mathematics and Statistics* 5(4):349–380.
26. Brenowitz ND, Bretherton CS (2018) Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters* 45(12):6289–6298.
27. Pelt DM, Sethian JA (2018) A mixed-scale dense convolutional neural network for image analysis. *Proceedings of the National Academy of Sciences* 115(2):254–259.
28. Majda AJ, Wang X (2006) *Nonlinear dynamics and statistical theories for basic geophysical flows.* (Cambridge University Press).