nature computational science

Article

https://doi.org/10.1038/s43588-022-00376-0

Discovering and forecasting extreme events via active learning in neural operators

Received: 5 April 2022

Accepted: 14 November 2022

Published online: 19 December 2022

Check for updates

Ethan Pickering[1] ✉, Stephen Guth[1], George Em Karniadakis[2] & Themistoklis P. Sapsis[1] ✉

Extreme events in society and nature, such as pandemic spikes, rogue waves or structural failures, can have catastrophic consequences. Characterizing extremes is difficult, as they occur rarely, arise from seemingly benign conditions, and belong to complex and often unknown infinite-dimensional systems. Such challenges render attempts at characterizing them moot. We address each of these difficulties by combining output-weighted training schemes in Bayesian experimental design (BED) with an ensemble of deep neural operators. This model-agnostic framework pairs a BED scheme that actively selects data for quantifying extreme events with an ensemble of deep neural operators that approximate infinite-dimensional nonlinear operators. We show that not only does this framework outperform Gaussian processes, but that (1) shallow ensembles of just two members perform best; (2) extremes are uncovered regardless of the state of the initial data (that is, with or without extremes); (3) our method eliminates 'double-descent' phenomena; (4) the use of batches of suboptimal acquisition samples compared to step-by-step global optima does not hinder BED performance; and (5) Monte Carlo acquisition outperforms standard optimizers in high dimensions. Together, these conclusions form a scalable artificial intelligence (AI)-assisted experimental infrastructure that can efficiently infer and pinpoint critical situations across many domains, from physical to societal systems.

The grand challenge of predicting disasters remains an extremely difficult and unsolved problem[1]. Disasters, such as pandemic spikes, structural failures, wildfires or rogue waves (rare, giant waves that pose a danger to ships and offshore structures[2,3]), are uniquely challenging to quantify. This is because they are both rare and arise from an infinite set of physical conditions[4]. The proposition of predicting extremes is analogous to finding a catastrophic needle in an infinite-dimensional haystack. This calls for methods that can both discover extreme events and encode physical phenomena into their modeling strategy. We present a Bayesian-inspired experimental design (BED) approach, described in detail in Fig. 1a, that addresses both challenges by combining a probabilistic 'discovery' algorithm[5,6] with a deep neural operator (DNO) designed to approximate physical systems[7].

The discovery of extremes is often simplified by distilling complex systems to their governing input variables and relevant output variables. Within this interpretation, quantification of extremes has historically taken the form of importance sampling, which uses a biasing distribution to identify regions of the input space that exhibit extreme values[8,9]. Unfortunately, these techniques often require additional and challenging considerations for accurate results[10–12]; they are also static, lacking an ability to adjust to new information gained through experiments. Active learning, specifically BED, provides a dynamic approach that learns from acquired data before selecting new and intriguing input–output data.

Active learning (AL) refers to a broad class of sequential sampling techniques for assembling efficient training datasets. AL has been applied with neural networks (NNs) in several fields, predominantly

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Division of Applied Mathematics, Brown University, Providence, RI, USA. ✉e-mail: pickering@mit.edu; sapsis@mit.edu

https://doi.org/10.1038/s43588-022-00376-0

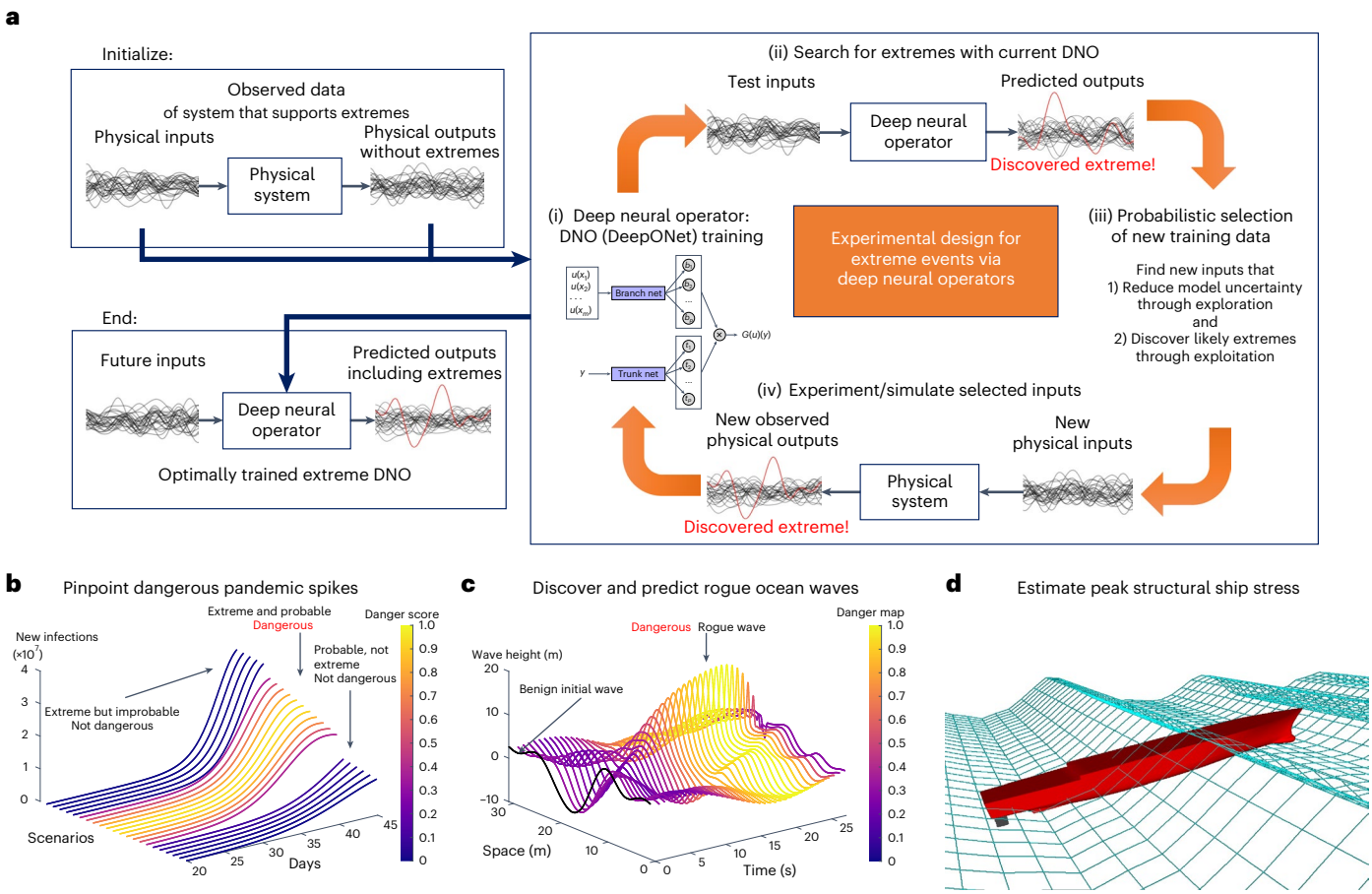

Fig. 1 | Active learning of extreme events in society and nature, from pandemic spikes to rogue waves, to structural ship failures. a, Efficient and robust DNO + BED framework for discovering and quantifying extremes. An overview of the proposed Bayesian experimental design framework with DNOs and rare-event acquisition functions for discovering extremes. *Initialize* with a set of observed physical input–output pairs, retained in their functional form. (i) Pass functions to an ensemble of DNOs to learn sparse representations of the underlying system. (ii) Perform a fast Monte Carlo search of the DNO functional space for extremes. (iii) Compute statistics over a Monte Carlo ensemble and select new input functions that both explore and exploit the space for extremes. (iv) Evaluate proposed inputs on the underlying experiment or simulation, record outputs (QoIs) and pass to (i). Repeat (i)–(iv) until statistics are converged or resources are depleted. *End* with an optimally trained DNO that supports prediction of extreme events. **b**–**d**, Inference of diverse extreme phenomena, from pandemic spikes to rogue ocean waves to large structural stress events on ships. Our framework pinpoints the most dangerous (that is, probable and extreme) pandemic scenarios (**b**), via realizations of stochastic infection rates, discovers rogue waves (**c**), and efficiently estimates large structural stress events for reliable ship design (**d**). In **a**, the most dangerous pandemic scenarios are pinpointed by inferring the number of new infections at time $T = 45$ days for a plurality of time-varying infection rate hypotheses. See Fig. 2 for more details. In **b**, rogue waves are discovered and quantified for future prediction by uncovering the probable wave conditions that nonlinearly interact in time to emit rogue waves over three times their original size. We show one example of this phenomena here and refer to Fig. 4 for more details on discovering these waves. In **c**, the statistics of peak stress govern fatigue lifetimes; with our approach we can efficiently estimate how unique ship designs structurally react to stochastic ocean waves to inform reliable and safe ship design. See Fig. 5 for the stress state related to this graphic.

in classification tasks such as image recognition[13], text recognition[14] or object detection[15] (see ref. [16] for a survey of similar AL applications), with less attention in the literature focused on regression of physical processes, let alone rare events. Although there are some exceptions for AL in rare-event quantification, such as combining deep NNs (DNN)[17] or other surrogate models[18] with weighted importance sampling for structural reliability analyses, neither leverage uncertainty to ensure the input space has been adequately explored. On the other hand, techniques employing BED and uncertainty predictions via Kriging or Gaussian process (GP) regression[5,19] perform well, but cannot be applied to infinite-dimensional systems or scale to large-training sets. This necessitates a solution that can (1) accurately generalize to infinite-dimensional systems and easily scale with data size, (2) emit uncertainty estimates and (3) apply appropriately defined acquisition functions for selecting extreme data.

DNOs, such as DeepONet[7], are built specifically for handling infinite-dimensional systems and provide the ideal surrogate model for characterizing extremes. Unlike other machine learning (ML) approaches, such as GPs, which map parameterizations of physical phenomena, DNOs directly map physical, infinite-dimensional functions to physical, infinite-dimensional functions. This leads to drastic improvements in generalization to unseen data in high dimensions. Additionally, the NN backbone of DNOs mean they are intrinsically amenable to big data, unlike GPs, which scale as the third power of data size[20,21]. However, the utility of DNOs for Bayesian experimental design is an open question, as DNOs do not explicitly provide a measure of uncertainty. We propose and show the efficacy of using an ensemble of DNOs for uncertainty quantification and BED. Although much of the literature is skeptical of the generality of ensembles to provide uncertainty estimates, recent viewpoints[22], notably ref. [23], have argued that DNN ensembles provide a very good approximation of the posterior. Our results support this perspective.

Appropriately defined acquisition functions for uncovering extreme behavior are just as critical as the chosen surrogate model. Recently, Sapsis and Blanchard[5,24], in concert with several other

https://doi.org/10.1038/s43588-022-00376-0

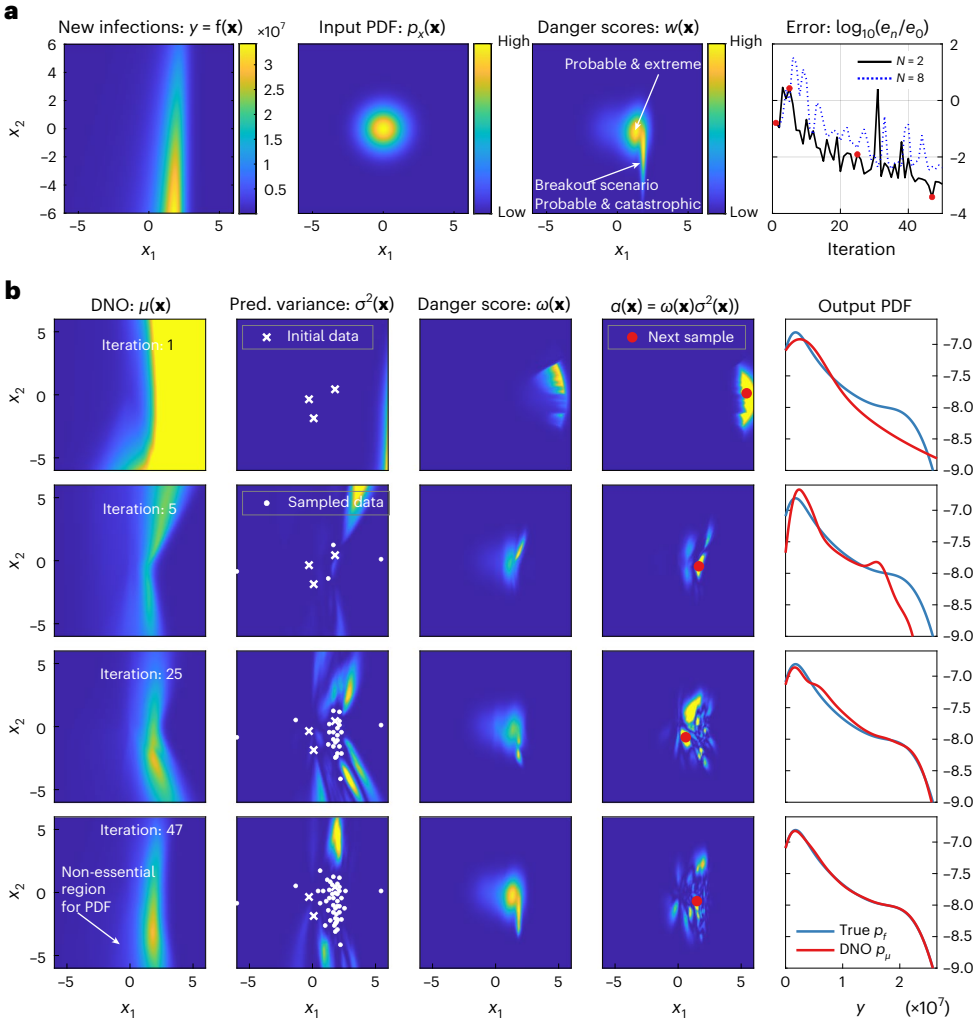

Fig. 2 | Nailing the tail, accurate PDF and danger score convergence in 50 samples. a, From left to right, the full deterministic response of new infections, $G(\mathbf{x})$, with respect to the two random parameters, x_1 and x_2, the probability distribution of the random parameters x_1, x_2, the underlying danger scores, $w(\mathbf{x})$ (with two regions of danger, one probable and high magnitude (that is, extreme) and one that is probable with catastrophically high magnitudes, constituting a breakout event), and the $\log_{10}$ of the normalized ($e_0 = 10^7$) log-PDF error (equation (1)), for the experiment performed in **b** using $N = 2$. The red circles indicate the iterations shown in **b** as well as additional results for a case with $N = 8$ ensembles. **b**, One experiment of the 2D stochastic SIR model using three initial samples and iterated 50 times. The rows represent iteration numbers 1, 5, 25 and 47 from top to bottom, respectively. The columns, from left to right, are (1) the DNO approximation of the objective function, $\mu(\mathbf{x})$, given the training samples (initial + acquired), (2) the samples acquired (where initial samples are shown as white crosses and acquired samples as white dots) in the 2D parameter space and the predictive variance $\sigma^2(\mathbf{x})$, (3) the calculated danger scores $w(\mathbf{x})$ and (4) the acquisition values $a(\mathbf{x})$, with the next acquisition sample denoted by a red circle, and (5) the DNO approximated and true output PDFs. Iteration 47, first column, identifies a high-magnitude infection region ignored by the algorithm, due to low input probability. Animation links presenting 100 iterations are provided for $N = 2$ (https://github.com/ethan-pickering/dnosearch_nature_cs_data/blob/main/movies/movieN2_Seed3.avi) and $N = 8$ (https://github.com/ethan-pickering/dnosearch_nature_cs_data/blob/main/movies/movieN8_Seed3.avi).

works[6,25], introduced a class of probabilistic acquisition functions specifically designed for quantifying extreme events under asymptotically optimal conditions[26]. By combining the statistics of the input space along with statistics deduced from the surrogate model, the method can account for the importance of the output relative to the input. This approach substantially reduces the number of input samples required to characterize extreme phenomena.

The main contribution of this work is a scalable, model-agnostic, Bayesian-inspired DNO–BED framework with extreme acquisition functions that efficiently learns to discover and forecast extreme events, Fig. 1a. This AI-assisted framework comes with several favorable properties, such as improved data acquisition efficiency, computational tractability, robustness and ease of implementation. The most important are summarized below:

1. The DNO framework is consistently more efficient than GP approaches, especially as dimensionality increases.
2. Shallow ensembles of just *two* members perform best, greatly reducing the training cost of ensemble DNOs.
3. The use of batches of suboptimal acquisition samples compared to step-by-step global optima does not hinder DNO–BED performance, permitting parallel experimentation in real-life applications.
4. Extremes are uncovered, regardless of the state of the initial data (that is, with or without extremes).
5. The method is observed to eliminate sample-wise 'double-descent' phenomena.

Equipped with both extreme acquisition functions and an ensemble of DNOs, our study demonstrates the above contributions

https://doi.org/10.1038/s43588-022-00376-0

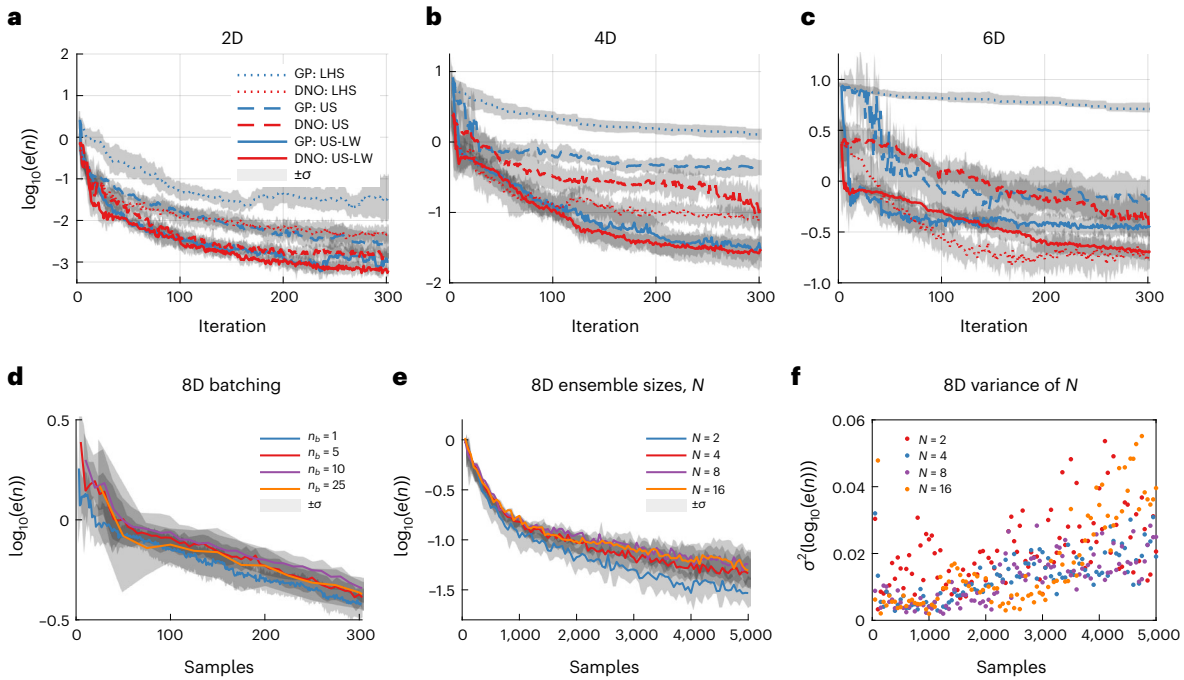

Fig. 3 | Accelerated convergence with DNOs plus extreme acquisition functions regardless of dimensionality and parallel acquisition and shallow ensembles bring computational efficiency without performance loss. a–**c**, Median log-PDF error, equation (1), of ten independent and randomly initialized experiments, where shaded regions denote 1 s.d. for each case (that is, $\pm\sigma(\log_{10}(e(n)))$), considering the three acquisition functions (LHS, US, US-LW) and GP and DNO surrogate models. **d**–**f**, Errors related to different batch sizes, n_b, per iteration (**d**), performance differences relative to the DNO ensemble size (100 iterations at batch size $n_b = 50$) (**e**) and the variance of the log-PDF error shown in **e** (**f**). In **a**–**f**, 2D, 4D and 6D ICs (**a**–**c**) and results from 8D ICs using the DNO surrogate model with US-LW (**d**–**f**) are shown.

by testing three classes of representative high-dimensional stochastic nonlinear system to discover extreme rogue waves, pinpoint dangerous pandemic scenarios, and efficiently estimate structural stresses to inform ship design (Fig. 1b–d).

Results

Our goal is to accurately quantify the probability distribution function (PDF), $p_y(y)$, of a stochastic quantity of interest (QoI), y. The variable y results from an observed random variable input, $\mathbf{x}$, that is transformed by the underlying system or map, $y = G(\mathbf{x})$. Although the statistics of y can be found through Monte Carlo sampling of the system, doing so is extremely inefficient. Instead, we aim to estimate an approximate map, $\tilde{G}$, via a surrogate model (for example, DNO or GP regression) trained on n observed data pairs $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$. The generated surrogate model may then be sampled over $\mathbf{x}$ at a substantially greater efficiency than the original system responsible for the observed (training) data, $\mathcal{D}$. Using the Bayesian surrogate model approach, we estimate the mean model, $\mu(\mathbf{x})$, by considering the mean over an N-ensemble of trained DNOs (each initialized using random weights). Note that, for the case of GPs, this mean can be calculated in closed form using standard expressions from GP regression.

Having the estimated mean model for the QoI, $\mu(\mathbf{x})$ and the PDF of random input $\mathbf{x}$, $p_{\mathbf{x}}(\mathbf{x})$, we can estimate the PDF for the QoI, $p_\mu(y)$, via a weighted kernel density estimator (KDE). This is done by computing a large number of samples distributed over the input space with latin hypercube sampling (LHS), $\mathbf{x}_j$, evaluating them with the surrogate map, $\tilde{y}_j = \mu(\mathbf{x}_j)$, determining their probability of occurrence, $\alpha_j = p_{\mathbf{x}}(\mathbf{x}_j)$, and estimating the PDF, $p_\mu(y) = \mathrm{KDE}(\mathrm{data} = \tilde{y}_j, \mathrm{weights} = \alpha_j)$, using standard Gaussian KDE implementations. To emphasize the role of rare and extreme events, we assess the surrogate approximation by the error metric

$$e = \int |\log_{10} p_\mu(y) - \log_{10} p_y(y)|\, \mathrm{d}y, \tag{1}$$

where the integral is computed over a finite domain for the QoI, extended over the values we are interested in describing statistically.

Approximating the underlying map with a surrogate model may require substantial data depending on the complexity and dimension of the input space. To reduce the amount of necessary training data, we combine rare event statistics and Bayesian experimental design approaches to uncover the most critical data for training the surrogate model with surrogate map, $\tilde{G}$, that ultimately produces an accurate PDF, including the tails, of the QoI, y.

Figure 1b–d presents the key implications of our results, diagnosing the most dangerous future pandemic scenarios (Fig. 1b; realizations of stochastic infection rates), discovering seemingly benign waves that lead to dangerous rogue waves (Fig. 1c), and identifying waves that lead to large structural stresses (Fig. 1d). In each case, different scenarios are tied to a 'danger score' or likelihood ratio, as proposed by Blanchard and Sapsis[5]:

$$w(\mathbf{x}) = \frac{p_{\mathbf{x}}(\mathbf{x})}{p_\mu(\mu(\mathbf{x}))}. \tag{2}$$

The likelihood ratio appropriately balances events that are probable and those that are extreme, so it provides a danger score for any given event. As denoted for the pandemic model in Fig. 1b, small danger scores are attributed to events that are either implausible or not extreme, whereas large danger scores relate to those that are both probable and extreme. However, any system's danger score requires knowledge of the true PDF, p_y, which is generally unknown and must be learned. Our approach efficiently learns this underlying distribution through dynamic application of the danger score with Bayesian experimental design and DNOs.

Pinpointing dangerous pandemic scenarios

Figure 2a,b demonstrate how the proposed active learning framework leverages dynamically updating danger scores and predictive variances

https://doi.org/10.1038/s43588-022-00376-0

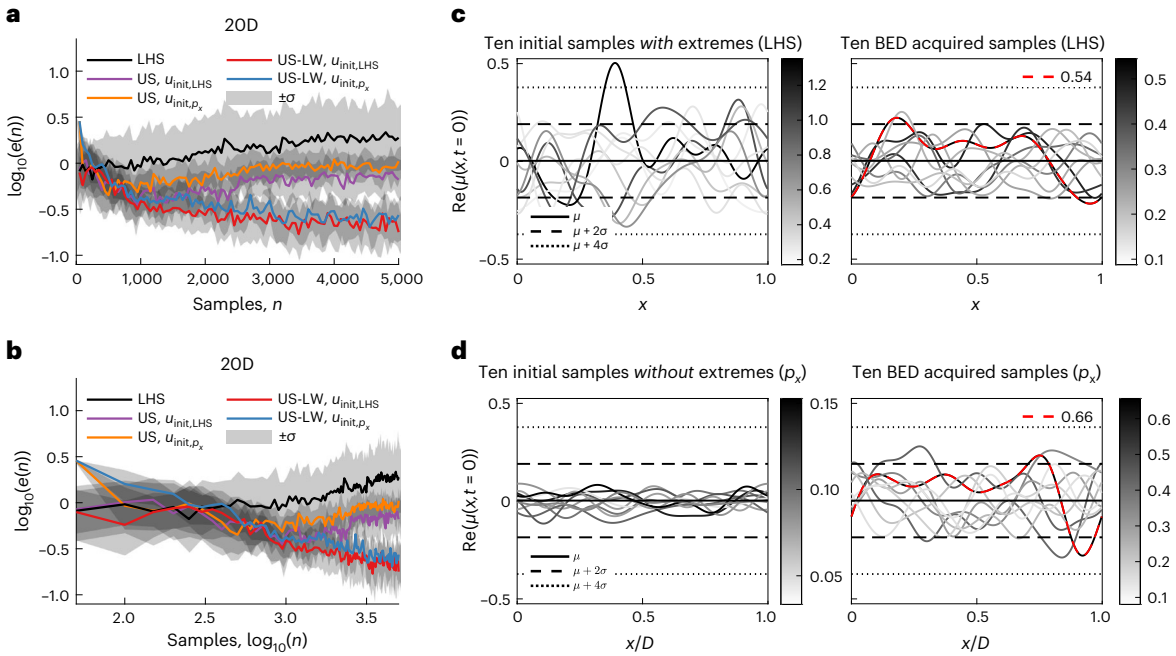

Fig. 4 | Robustness to dimensionality and initial data, with or without extremes. a, The median log-PDF errors, where shaded regions denote 1 s.d. for each case (that is, $\pm\sigma(\log_{10}(e(n)))$), for three acquisition functions, LHS, US and US-LW, for the 20D problem, with initial samples, u_{init}, sampled via LHS and the prior p_x for US and US-LW. **b**, The same results in log–log form. **c**, The real value of ten complex LHS initial input functions $u(x, t = 0)$, colored with their corresponding QoI, $||\text{Re}(u(x, t = T))||_\infty$ (left) and ten DNO–BED acquired functions, via US-LW at iteration 97 (right). **d**, Ten initial input functions sampled from the prior distribution p_x (left) and the ten acquired functions at iteration 74 (right).

to efficiently sample the underlying system and learn the PDF of infections for a stochastic pandemic model. The pandemic model is the simple 'susceptible, infected, recovered' (SIR) model, proposed by Kermack and McKendrick[27] and popularized by Anderson and May[28], with a two-dimensional stochastic infection rate (see Supplementary Section 1 for details). We assess success as the log-PDF error (equation (1); Fig. 2a, last column) between the true output distribution and the approximated distribution from the trained DNO (Fig. 2b, last column). See the section QoI and log-PDF error metrics for details on computing the log-PDF error metric.

Our framework quickly identifies the key regions of dynamical relevance and accurately recovers the important properties of the underlying system, despite an initialization of only three data samples in the parameter space. Using an ensemble of just two DNOs (Supplementary sections 5 and 6 describe DNO implementation and the section Ensemble of neural networks for uncertainty quantification describes our application of ensemble methods), the algorithm iteratively provides an estimation of the underlying map (for computing p_μ), predictive variance $\sigma^2(\mathbf{x})$ and a danger score $w(\mathbf{x})$. Together, the danger score and predictive variance create the likelihood-weighted uncertainty sampling (US-LW) acquisition function, $a(\mathbf{x}) = w(\mathbf{x})\sigma^2(\mathbf{x})$ (ref. [5]), which identifies the sample within the parameter space with the greatest potential for learning the true output PDF. With the addition of each point, all fields dynamically change and bring the true and approximated output PDFs within greater agreement. By iteration 50, the danger scores have converged and the approximated output PDF has an error of less than 10^{-3}. It is from this final danger score map that we derive the pandemic scenarios of Fig. 1b. This plot of danger scores includes two regions, one with probable and high-magnitude pandemic spike scenarios and another breakout scenario with catastrophic consequences (an exceptionally high magnitude of infections). Each region is annotated in Fig. 2a, third column. Additionally, the last column of Fig. 2a shows that increasing the ensemble size to $N = 8$ provides little to no advantage.

The critical aspect of this approach is the algorithm's reduction of a large parameter space to local regions of danger. Only regions that provide critical contributions to the output PDF are considered. Iteration 47 (last row, first column, in Fig. 2b) underscores this behavior. The algorithm has accurately reconstructed the output PDF, but, by juxtaposing iteration 47 with Fig. 2a, we see that it has ignored a region where infections are of high magnitude located at $x_1 \approx 1.5$ and $x_2 = [-5, -6]$. This region, as well as the remaining unexplored regions, provides negligible information about the QoI and is neglected by the acquisition function. This property is crucial for all systems where resources for experiments or simulation are limited or costly, and it permits a substantial reduction in training/acquired data as system complexity and dimensionality increases.

Discovering and predicting rogue waves

We now train a surrogate model for rogue-wave prediction by actively discovering the probable initial conditions (ICs), or precursors, responsible for such phenomena. Here we present a proof of concept with a dispersive nonlinear wave model proposed by Majda, McLaughlin and Tabak (MMT)[29,30] for one-dimensional (1D) wave turbulence. The same model has also been used as a prototype system to model rogue waves[31–34] (Supplementary Section 2).

We seek to map initially observed waves, $u(x, t = 0)$, where x is the spatial variable and t is time, to the QoI: the future spatial maximum $G(\mathbf{x}) = ||\text{Re}(u(x, t = T; \mathbf{x}))||_\infty$, where T is a prescribed prediction horizon. We note that MMT has complex solutions and therefore the ICs are also complex-valued. In a real application, other quantities, such as the short time derivative, would accompany the initial condition, instead of an imaginary component, to provide wave speed. We now investigate this complex and highly nonlinear problem by systematically scaling the dimensionality and expanding to larger datasets, where GPs begin to fail.

The DNO–BED framework, using the US-LW acquisition function ($a(\mathbf{x}) = w(\mathbf{x})\sigma^2(\mathbf{x})$), efficiently minimizes the error between the approximated and true PDF of the QoI when compared to GPs (as detailed in

https://doi.org/10.1038/s43588-022-00376-0

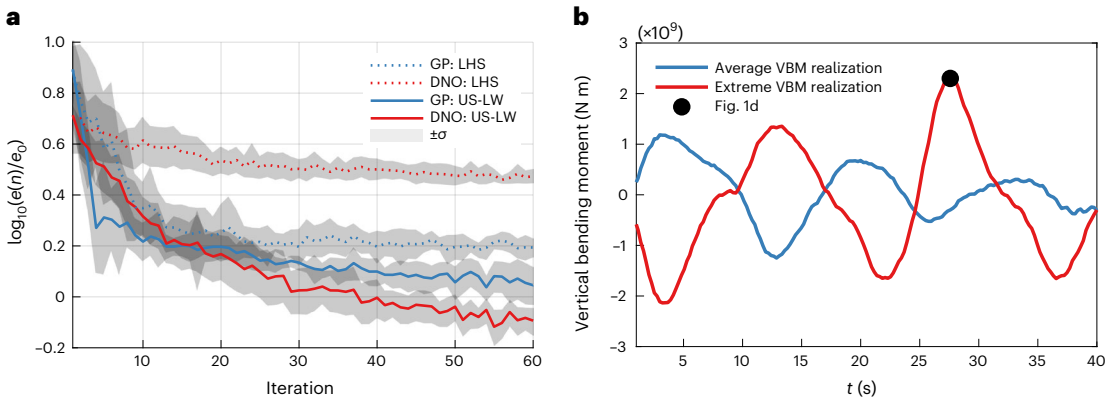

Fig. 5 | Efficient learning of fatigue statistics for ship design. a, Median-normalized ($e_0 = 5 \times 10^8$) log-PDF errors for ten experiments of VBM statistics for GP and DNO using LHS and US-LW, where the shaded regions denote 1 s.d. (that is, $\pm\sigma(\log_{10}(e(n)))$). **b**, Two representative VBM curves of average (blue) and extreme (red) loads on the ship from stochastic ocean waves. The black dot indicates the instantaneous VBM sustained by the ship in Fig. 1.

Supplementary Section 4) or other common BED sampling strategies, such as US, $a(\mathbf{x}) = \sigma^2(\mathbf{x})$, and LHS, as shown in Fig. 3a–c. This is highlighted by the dimensionality (that is, complexity) of the parameterized ICs increasing from 2D to 6D, where GPs begin to break down. Details on dimensionality are provided in Supplementary Section 2.

The DNO–BED framework using US-LW acquisition functions also brings favorable results for reducing computational cost in high dimensions, as both batching samples via parallel selection of multiple acquisition points (Fig. 3d) and the use of shallow ensembles of only $N = 2$ members (Fig. 3e) perform without loss. Computed at 8D, Fig. 3d shows that, regardless of choosing $n_b = 1$, 5, 10 or even 25 samples per iteration, over 300 samples, does not result in a loss of performance with our framework. Furthermore, we describe in Supplementary Section 7 that the batched samples come from several regions of local optima and the use of Monte Carlo methods is substantially more efficient for identifying these acquisition samples than standard Python optimizers. See the section Experiment batching for details on batching implementation. These observations are a critical result for scaling our framework, as more complex systems inevitably will require more data.

Unexpectedly, not only do just two ensemble members, $N = 2$, perform well, they consistently outperform larger ensembles, $N > 2$, in Fig. 3e (where 100 iterations of batch size $n_b = 50$ are applied to the 8D case). This result appears to disagree with the natural hypothesis that a larger set of ensembles would provide uncertainty estimates with greater fidelity, leading to better performance of our sequential search methods. Clearly, the latter is not the case from our results, yet neither can it be that $N = 2$ ensembles provide a predictive variance of greater fidelity than $N = 16$. Figure 3f permits both concepts to be true. It shows that the greater the ensemble size, the smaller the variance between the error trajectories of independent experiments. This observation agrees with the idea that larger ensembles lead to a higher-fidelity predictive variance, but that greater fidelity leads to consistency rather than performance for this sequential search technique. We believe that using small N imposes a greedy search, in a similar fashion to Thompson sampling[35]. Regardless, the consistent observation that $N = 2$ is not only viable, but perhaps preferable, substantially minimizes computational costs for ensemble approaches.

Rogue-wave discovery in 20D. Equipped with the computational advantages of $N = 2$ ensemble members and large batch sizes ($n_b = 50$), Fig. 4a,b shows that, even at 20D, our approach can recover the QoI PDF. The other acquisition functions not only perform poorly, but perform worse as more data are acquired. This observation appears to be related to the phenomenon known as sample-wise 'double descent', and many researchers have observed this behavior throughout ML procedures, from classification to regression problems[36]. Sample-wise double descent is associated with instabilities in the surrogate model, a product of overfitting. More data and greater complexity results in an over-parameterization of the provided data. Temporarily, this leads to inferior generalization until providing the surrogate model with sufficiently larger datasets.

The proposed acquisition function clearly avoids log-PDF error double descent for the 20D problem. Although we do not explicitly detail the mechanism behind this observation here, we refer to a parallel study by Pickering and Sapsis[37] on this exact problem in 8D (showing elimination of both mean square error and log-PDF error double descent, as well as other examples using GP surrogate models) and briefly outline why the acquisition brings this beneficial behavior. As discussed in ref. [37], double descent is eliminated by only selecting data that critically contribute to the observed dynamics of the system. Unlike US-LW, the data chosen by LHS and US methods are not inherently important to recovering the true PDF and therefore induce misleading complexity in the underlying regression task. This observation further underscores the value of our acquisition function, as it systematically prevents overfitting and unwarranted model complexity.

Additionally, we find that, regardless of the origin of the initial samples—with extremes (chosen by LHS) or without extremes (from the prior, p_x)—the method achieves similar error metrics in Fig. 4a,b. To elucidate the physical meaning of these results, Fig. 4c,d presents the physical manifestations of the 20D initial conditions, that is, the real component of $u_{\text{init}}(x, t = 0)$, for those that include extremes and those that do not, respectively. The horizontal lines denoting the mean and standard deviations (μ, σ) refer to the statistics of the normally distributed ICs, $u_{\text{init}}(x, t = 0)$, not the QoI. Instead, the QoI is observed after evolving the presented ICs by T, and the associated QoI value is represented by the color of the line. Focusing on Fig. 4c (left), several of the initial LHS ICs are already extreme, as portions of the ICs already exceed 3 s.d. (the common delimiter of extremes for normal distributions). Meanwhile, Fig. 4d (left) gives IC samples from the prior, which are nearly bounded by 1 s.d. Despite these clearly different sets of initial ICs, the algorithm samples similar ICs in the right plots of Fig. 4c,d, and achieves similar error metrics in Fig. 4a,b. These ICs are approximately bounded within $\pm 2\sigma$, meaning they are neither extreme (yet) nor common, but sit on the periphery of a dynamical instability that may lead to an extreme event. Therefore, the method is able to uncover the seemingly benign conditions that lead to extremes, regardless of the initial training set. Finally, with respect to the IC statistics, observed QoIs above ~0.3 constitute an extreme event at $t = T$.

https://doi.org/10.1038/s43588-022-00376-0

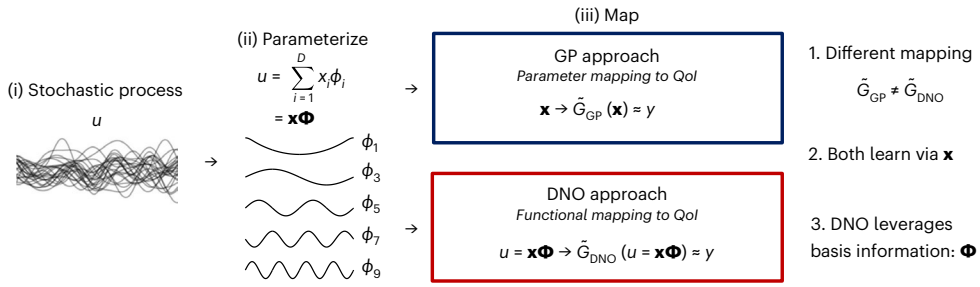

Fig. 6 | DNOs leverage functional information for mapping to the QoI. Both GPs and DNOs learn a stochastic process via a parameterization of a stochastic process, but differ in that GPs map the parameter space to the QoI, and DNOs map the functional realization of the parameter space to the QoI.

Efficient estimation of structural fatigue for ship design

Uncovering the statistical signature of a marine vessel's vertical bending moment (VBM) at the midship section is critical to estimating fatigue lifetimes. Caused by cyclic hydrostatic pressure forces and the slamming of a ship's bow into oncoming waves[38–40], large VBMs increase the potential for microfracture nucleation and propagation[41–43]. Just as in our previous examples, these forces are stochastic processes that result in unique VBM statistics for each unique ship design (such as the Office of Naval Research topsides flare variant studied here) via an underlying nonlinear operator. Performing either tow-tank or numerical experiments is expensive and time-consuming, limiting the potential for optimal structural design. Here we apply our method to a proprietary code, LAMP (Large Amplitude Motion Program, v4.0.9, May 2019)[44], to calculate the expensive forward problem of a specific VBM response to a specific wave episode for informing ship design.

Applying the framework to a 10D subspace of finite-time ocean-wave episodes impacting the ship modeled in LAMP, we again find that the US-LW DNO is most efficient in learning the VBM (QoI) statistics when compared with the other methods shown in Fig. 5a. To highlight the relative ease of implementation of the DNO-based approach, we emphasize that the GPs were very carefully tuned (optimization of hyperparameters using additional simulations, which took a substantial amount of time and effort), in contrast to the minimal tuning efforts required for DNOs. Figure 5b also presents two representative realizations, one average and one extreme, of the VBM time series felt by the ship through a set of ocean waves. The highlighted point on the extreme realization at $t = 27$ represents the instantaneous VBM load that the ship encounters during an instant as presented earlier in this Article (Fig. 1d).

Discussion

Although we have only shown the ability of the DNO–BED framework to efficiently learn extreme statistics for three example problems, prototype rogue waves, pandemic spikes and ship fatigue, this approach is general for learning any stochastic nonlinear system with extreme events. However, we caution readers that our results are empirical and do not come with robust guarantees. This is compounded by the non-analytical nature of NNs, which limit hope that guarantees may ever be found. Thus, implementation must be performed carefully and systematically. Training the DNOs is not trivial and requires finesse, while the acquisition function, although not observed here, could fall prey to unforeseen pathological cases. To help readers implement the method, Supplementary Section 8 provides several tips as well as general implementation concerns that we either experienced or recognized during implementation.

The framework also provides modularity for the chosen surrogate models and acquisition functions, as long as the parameter and function spaces are appropriately defined. Critically, we separate the two, as shown in Fig. 6. In GP implementations, the parameter space of a stochastic process is used for both regression and searching. Instead, our implementation of DNOs performs regression in the functional space, leveraging the typically disregarded basis functions associated with the parameterization, and the search algorithm is performed in the parameter space via a forward DNO coupling with the associated basis functions. It is this maintenance of the functional representation that provides improved generalization across the parameter space.

As such, any arbitrary neural architecture leveraging this distinction may be implemented, such as standard feedforward NNs, Fourier neural operators[45], convolutional NNs, recurrent NNs, long-short term memory, among others. In fact, this work did not explicitly investigate the full value of DeepONet for operator learning for BED. Instead, only the standard feedforward branch NN was used. Using the complete operator machinery only requires adjusting the parameters related to the operator and trunk. Here, we kept these parameters constant. Finally, while we focus on Bayesian experimental design, slight adjustments to the choice of acquisition function allow for Bayesian optimization tasks[46] or for approaching other metrics of interest (for example, mean squared error).

In conclusion, we believe we have demonstrated an equation-agnostic framework that (1) efficiently discovers extreme events, (2) is computationally tractable and (3) can be implemented straightforwardly on any stochastic input–output system. This creates a unique opportunity for experimentalists and computationalists alike to investigate and quantify their stochastic systems with respect to extreme behavior, whether that behavior originates from societal or physical systems or has beneficial or catastrophic consequences.

Methods

Our approach to Bayesian experimental design, detailed in Fig. 1a, consists of two critical components, data selection criteria and the surrogate model. Supplementary Algorithm 1 formalizes the iterative steps taken in Fig. 1a for efficiently training a surrogate model with minimal data selection.

Surrogate models

We test two surrogate models in the BED framework, GPs (as the benchmark case) and DNOs (specifically DeepONet) for greater scaling and generalization performance. Although both GPs and DNOs provide the same role in the framework, their implementation is fundamentally different (Fig. 6 provides a visual representation).

Whereas GPs are used to approximate the map of random parameter inputs, $\mathbf{x} \in \mathbb{R}^{1\times D}$ (D is the parameterization dimension), to a QoI, $\mathbf{x} \mapsto \tilde{G}_{GP}(\mathbf{x})$, we use DNOs to approximate the map of the parameterized input function, $u = \mathbf{x}\boldsymbol{\Phi}$, to the QoI, $u \mapsto \tilde{G}_{DNO}(u)$. In the case of the MMT problem, the input function varies spatially over x as $u(x) = \mathbf{x}\boldsymbol{\Phi}(x)$, that is, the scalar product between $\mathbf{x}$ and $\boldsymbol{\Phi}(x)$, where $\boldsymbol{\Phi}(x) \in \mathbb{R}^{D\times n_x}$ is the parameterization basis over n_x spatial points, to an output function $G(u(x) = \mathbf{x}\boldsymbol{\Phi}(x))$. In both cases—GP and DNO—$\mathbf{x}$ is the only independent variable and is maintained to provide a means of searching for extremes in the parameter space. However, the proposed regression task for

DNOs ensures that they perform their mapping in functional, or physical, space, rather than the parameter space. This distinction is foundational for our approach and success with DNOs for BED.

We provide details for GPs and DeepONet in Supplementary 4 and 5, respectively, and only discuss our approach for quantifying the predictive variance, $\sigma^2(\mathbf{x})$, for DNOs using ensembles.

Ensemble of neural networks for uncertainty quantification. Although NN architectures are attractive for approximating nonlinear regression tasks, their complexity rids them of analytical expressions. This does not allow for a traditional Bayesian treatment of uncertainty in the underlying surrogate model—a key property present for GP regression. Knowledge of the uncertainty of a surrogate model allows one to target model deficiencies as seen in the parameter space. This means that choosing an appropriate method for quantifying the uncertainty is a crucial and key component to active learning or BED. There are several techniques for quantifying uncertainty in NNs; we provide a brief description of these in Supplementary Section 5.1, focusing only on ensemble methods.

Ensemble approaches have been used extensively throughout the literature[47,48] and, despite their improved results for identifying the underlying tasks at hand[49], their utility for quantifying uncertainty in a model remains a topic of debate. There are several approaches for creating ensembles. These include random weight initialization[48], different network architectures (including activation functions), data shuffling, data augmentation, bagging, bootstrapping and snapshot ensembles[50–52], among others. Here we employ random weight initialization, a technique found to perform similarly or better than Bayesian NN approaches (Monte Carlo dropout and probablistic backpropagation) for evaluation accuracy and out-of-distribution detection for both classification and regression tasks[48]. As stated earlier, much of the field is skeptical of the generality of ensembles to provide rigorous uncertainty estimates. However, recent studies, such as that in ref. [22] and specifically ref. [23], have argued that DNN and DNO ensembles provide reasonable, if not better, approximations of the posterior. Finally, the straightforward implementation of the randomly initialized weights motivates our choice, as it makes the adoption of these techniques far more probable.

We train N randomly weight-initialized DNO models, each denoted as $\tilde{G}_n$, that find the associated solution field y for functional inputs u and operator parameters z (that is, components that change the underlying operator, such as exponents or operator coefficients). This allows us to then determine the pointwise variance of the models as

$$\sigma^2(u, z) = \frac{1}{(N-1)} \sum_{n=1}^{N} \left(\tilde{G}_n(u)(z) - \overline{\tilde{G}(u)(z)}\right)^2, \tag{3}$$

where $\overline{\tilde{G}(u)(z)}$ is the mean solution of the model ensemble. Finally, we must adjust the above representation to match the description for BED. In the case of traditional BED and GPs, the input parameters, $\mathbf{x}$, represent the union of two sets of parameters, $\mathbf{x}_u$ and $\mathbf{x}_z$. The parameters $\mathbf{x}_u$ typically represent random variables applied to a set of functions that represent a decomposition of a random function $u = \mathbf{x}_u \mathbf{\Phi}(x_1, ..., x_m)$, where $x_1, ..., x_m$ are discrete function locations (spatial, temporal or both), and $\mathbf{x}_z = z$ represents non-functional parameters of the operator. Thus, the DNO description for uncertainty quantification may be recast as

$$\sigma^2(\mathbf{x}) = \sigma(\mathbf{x}_u \cup \mathbf{x}_z) = \frac{1}{(N-1)} \sum_{n=1}^{N} \left(\tilde{G}_n\left(\mathbf{x}_u \mathbf{\Phi}(x_1, ..., x_m)\right)(\mathbf{x}_z) - \overline{\tilde{G}\left(\mathbf{x}_u \mathbf{\Phi}(x_1, ..., x_m)\right)(\mathbf{x}_z)}\right)^2. \tag{4}$$

Data selection and acquisition functions

The acquisition function is the key component of the sequential search algorithm, as it guides Supplementary Algorithm 1 in exploring the input/parameter space and determines samples at which the objective function is to be queried. Because of the lack of a closed analytical form of DNOs, we only consider two acquisition functions as used previously with GPs on several test cases in ref. [5]. The two functions we are interested in are the commonly used US and the output-weighted US, proposed in ref. [5] and shown in ref. [26], to guarantee optimal convergence in the context of GP regression. The difference here is that we apply them explicitly to DNOs (DeepONet) and present the advantages of DNOs compared to GPs as we apply them to a complex and high-dimensional problem.

Uncertainty sampling. US is one of the most broadly used active sampling techniques and identifies the sample where the predictive variance is the greatest:

$$a_{\mathrm{US}}(\mathbf{x}) = \sigma^2(\mathbf{x}). \tag{5}$$

US, also known as the active learning-MacKay (ALC) algorithm[53], imposes a sequential search that evenly distributes uncertainty over the input space as it gains data. The popularity of US is due to three qualities: ease of implementation, inexpensive evaluation (for small datasets with GPs) and analytic gradients, the last of which permits the use of gradient-based optimizers.

Likelihood-weighted acquisition functions. There are several 'extreme event' LW acquisition functions that could be explored, as proposed in ref. [6], but we elect to only test the US-LW acquisition because of its simplicity in implementation. For US-LW, we augment the US sampling acquisition function with the previously described danger scores to give

$$a_{\mathrm{US-LW}}(\mathbf{x}) = w(\mathbf{x})\sigma^2(\mathbf{x}), \tag{6}$$

such that both highly uncertain and high-magnitude regions are sampled.

To compute $w(\mathbf{x})$, we note that the approximated output PDF, $p_\mu(\mu)$ is approximated via a kernel density estimator with $n = 10^6$ test points (10^7 for the 20D example). For the DNO cases, we chose to compute this with only the first ensemble member, $\mu = \tilde{G}_1$, to reduce computational costs. Similar to the ensemble results for $N = 2$, using only one ensemble member is akin to using Thompson sampling[35,54] and performs without reduction in performance.

QoI and log-PDF error metrics

To test the ability of the DNO and GP Bayesian-inspired sequential algorithms to quantify extremes, we define a QoI, or 'danger map', for the pandemic scenarios and rogue waves. The rogue-wave QoI is defined as

$$G(\mathbf{x}) = ||\mathrm{Re}(u(x, t = T; \mathbf{x}))||_\infty, \tag{7}$$

where $T = 20$, and the pandemic QoI is

$$G(\mathbf{x}) = I(t = T; \mathbf{x}), \tag{8}$$

where $T = 45$ days.

For each case, we then select 10^5 LHS test samples, $\mathbf{X} \in \mathbb{R}^{d \times 10^5}$, evaluate the true QoI at each, $\mathbf{y} = G(\mathbf{X}) \in \mathbb{R}^{1 \times 10^5}$, compute the probability of each sample, $\alpha = p_{\mathbf{x}}(\mathbf{X}) \in \mathbb{R}^{1 \times 10^5}$ and find the true PDF, $p_G(y) = \mathrm{KDE}(\mathrm{data} = \mathbf{y}, \mathrm{weights} = \boldsymbol{\alpha})$, using standard Gaussian KDE implementations (for example, scipy.stats.gaussian_kde). The approximated PDF is then found by replacing the true map with the surrogate map at each iteration.

Although the 10^5 test points provide sufficient samples for accurate PDF assessment through a KDE at all cases 10D and less, the absolute truth PDF for the 20D example is much more difficult. To attain a definitively converged PDF would require >10^7 samples, a

computationally infeasible quantity. Therefore, we emphasize that our truth metric is based on how well the approximated 10^5 test points reconstruct the KDE PDF given the same true 10^5 test points, rather than a definitive converged truth. This means that as long as the true behavior and the surrogate model are the same on these 10^5 LHSs, then the PDFs generated are identical everywhere. Although this may seem like a simple task, our results only sample a maximum of 5,000 training points (Fig. 4), two orders of magnitude less than the test set. At this size, only the DNOs with output-weighted sampling are unable to accurately regress to the test set. Therefore, this shows that the method is able to learn the underlying map at a substantially improved efficiency.

Finally, to determine whether the testing data appropriately identifies extremes, we compute the log-PDF error:

$$e(n) = \int |\log_{10} p_{\mu_n}(y) - \log_{10} p_G(y)| \, \mathrm{d}y, \tag{9}$$

where n is the iteration number.

For the LAMP problem, where the output measure (VBM) is a time series, the QoI and the error metric are slightly more complicated. The GP and DNO map $\mathbf{x}$ to $\mathbf{q}$ as

$$G(\mathbf{x}) = \mathbf{q}(\mathbf{x}), \tag{10}$$

and the VBM in time is recovered via

$$y(t; \mathbf{x}) = \sum_{i=1}^{12} q_i(\mathbf{x}) \phi_{q_i}(t), \tag{11}$$

where $\phi_{q_i}(t)$ are output basis vectors in time. We then concatenate realizations of $y(t; \mathbf{x})$ to form a single, long time series $Y(t)$. We drop explicit dependence of $Y(t)$ on $\mathbf{x}$, under ergodicity assumptions. Finally, our quantity of interest is the one-point time statistics of $Y(t)$. The ground-truth PDF $p_y(y)$, is computed using 3,000 Monte Carlo realizations, each over 1,800 time units, and the surrogate model approximation, $p_{\mu_n}(y)$, uses 10,000 LHS realization of 40 time units. Both PDFs are computed using standard unweighted KDE, and the log-PDF error is computed as in equation (9).

Monte Carlo optimization of acquisition functions

In our experiments, we consistently observe that acquisition samples found through optimizers using gradient descent are not globally optimal. Instead, Monte Carlo evaluation of the DNOs and GPs consistently find improved optima. This is chiefly because of the non-convexity of the acquisition function. We may recall the highly non-convex behavior of the 2D acquisition fields in Fig. 2b, even with as few as approximately five samples. This non-convex nature emits many local minima that require many initial search samples to provide confidence that the chosen optima are nearly global. As the optimizer progresses for each iteration, it must call the DNO or GP, whereas a Monte Carlo approach may efficiently evaluate all samples in one vector operation. This means that for the same computation time as the optimizers, Monte Carlo sampling may evaluate a substantially larger distribution of query samples and return acquisition samples with greater scores than that of the optimizer. In Supplementary Section 7 we show that acquisition scores found via Monte Carlo at 20D consistently outperform optimizers for similar computation times.

Although we choose to implement a Monte Carlo approach instead of off-the-shelf optimizers, there are probably several other optimization approaches that would prove to perform better in finding optimal sets of acquisition points. However, our contribution is focused on defining and implementing the acquisition function in DNOs for selecting the next experiment, and we leave further optimization of the DNO acquisition space for future work.

Experiment batching

As systems become more complex, additional experiments/data are required to reduce errors for higher-dimensional cases, as observed in Fig. 3. Considering that many experiments can be conducted in parallel, we ask whether choosing multiple local minima of the acquisition function presents marginally reduced performance than a purely sequential search. This is especially critical for situations where experimental time is more costly than additional set-ups (for example, protein or genetic design).

The purpose of batching is to find multiple regions of local optima of the acquisition function, rather than finding several optima in the same region. To impose this idea, we create a constraint that no acquisition sample may reside closer than a distance $r_{\min}$ to one another. We define $r_{\min}$ as a fraction of the maximum euclidean distance of the space being sampled:

$$r_{\min} = r_l \left(\sum_{d=1}^{D} (x_{d,+} - x_{d,-})^2 \right)^2, \tag{12}$$

where $x_{d,+}$ and $x_{d,-}$ are the maximum and minimum domain bounds of each parameter dimension d and r_l is the user-defined percentage. In this work we chose a static $r_l = 0.025$, but dynamic values based on the packing of the parameter space would be an intriguing direction for increasing the efficacy of this approach. Imposing this constraint requires an iterative processing of the acquisition scores, as detailed in Supplementary Algorithm 2. For the batching applied in this Article, each case uses a Monte Carlo querying of $n_q = 10^6$ points.

Data availability

All relevant data for reconstructing the results, including the LAMP dataset, are provided at dnosearch_nature_cs_data[55]. Additionally, all data, with exception of the LAMP data, may be computed from scratch using the code found in the dnosearch[56] GitHub repository. Source data are provided with this paper.

Code availability

Code pertaining to the sequential discovery algorithm of the SIR, MMT and LAMP problems is publicly available from the GitHub repository dnosearch[56]. The DeepONet code framework can be found within the deepxde package on GitHub. Code pertaining to the Large Amplitude Motions Program (LAMP) v4.0.9 (May 2019) is a proprietary code developed by Leidos (formerly SAIC). Additional product information about LAMP may be found by contacting Leidos at https://www.leidos.com/contact.

References

1. *Creating a Disaster Resilient America: Grand Challenges in Science and Technology* (National Academies Press, 2005).
2. Hansteen, O. E., Jostad, H. P. & Tjelta, T. I. Observed platform response to a "monster" wave. in *Field Measurements in Geomechanics* 73–86 (Taylor & Francis, 2003).
3. Gemmrich, J. & Cicon, L. Generation mechanism and prediction of an observed extreme rogue wave. *Sci. Rep.* **12**, 1718 (2022).
4. Sapsis, T. P. Statistics of extreme events in fluid flows and waves. *Annu. Rev. Fluid Mech.* **53**, 85–111 (2021).
5. Blanchard, A. & Sapsis, T. Output-weighted optimal sampling for Bayesian experimental design and uncertainty quantification. *SIAM/ASA J. Uncertain. Quantif.* **9**, 564–592 (2021).
6. Blanchard, A. & Sapsis, T. P. Bayesian optimization with output-weighted optimal sampling. *J. Comput. Phys.* **425**, 109901 (2021).
7. Lu, L., Jin, P., Pang, G., Zhang, Z. & Karniadakis, G. E. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nat. Mach. Intell.* **3**, 218–229 (2021).

8. Kahn, H. & Marshall, A. W. Methods of reducing sample size in Monte Carlo computations. *J. Op. Res. Soc. Am.* **1**, 263–278 (1953).
9. Shinozuka, M. Basic analysis of structural safety. *J. Struct. Eng.* **109**, 721–740 (1983).
10. Dematteis, G., Grafke, T. & Vanden-Eijnden, E. Extreme event quantification in dynamical systems with random components. *SIAM/ASA J. Uncertain. Quantif.* **7**, 1029–1059 (2019).
11. Uribe, F., Papaioannou, I., Marzouk, Y. M. & Straub, D. Cross-entropy-based importance sampling with failure-informed dimension reduction for rare event simulation. *SIAM/ASA J. Uncertain. Quantif.* **9**, 818–847 (2021).
12. Wahal, S. & Biros, G. BIMC: the Bayesian Inverse Monte Carlo method for goal-oriented uncertainty quantification. Part I. Preprint at https://arxiv.org/abs/1911.00619 (2019).
13. Gal, Y., Islam, R. & Ghahramani, Z. Deep Bayesian active learning with image data. In *Proc. International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 1183–1192 (PMLR, 2017).
14. Zhang, Y., Lease, M. & Wallace, B. Active discriminative text representation learning. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 31, 3386–3392 (AAAI, 2017).
15. Aghdam, H. H., Gonzalez-Garcia, A., van de Weijer, J. & López, A. M. Active learning for deep detection neural networks. In *Proc. IEEE/CVF International Conference on Computer Vision* 3672–3680 (IEEE, 2019).
16. Ren, P. et al. A survey of deep active learning. *ACM Comput. Surveys* **54**, 1–40 (2021).
17. Xiang, Z., Chen, J., Bao, Y. & Li, H. An active learning method combining deep neural network and weighted sampling for structural reliability analysis. *Mech. Syst. Signal Process.* **140**, 106684 (2020).
18. Ehre, M., Papaioannou, I., Sudret, B. & Straub, D. Sequential active learning of low-dimensional model representations for reliability analysis. *SIAM J. Sci. Comput.* **44**, B558–B584 (2022).
19. Echard, B., Gayton, N. & Lemaire, M. AK-MCS: an active learning reliability method combining Kriging and Monte Carlo simulation. *Struct. Safety* **33**, 145–154 (2011).
20. Snelson, E. & Ghahramani, Z. Sparse Gaussian processes using pseudo-inputs. *Adv. Neural Inf. Process. Syst.* **18**, 1257–1264 (2006).
21. Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. In *Proc. Artificial Intelligence and Statistics* (eds van Dyk, D. & Welling, M.) 567–574 (PMLR, 2009).
22. Pickering, E. & Sapsis, T. P. Structure and distribution metric for quantifying the quality of uncertainty: assessing Gaussian processes, deep neural nets and deep neural operators for regression. Preprint at https://arxiv.org/abs/2203.04515 (2022).
23. Wilson, A. G. & Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. In *Proc. 34th International Conference on Neural Information Processing Systems* (eds Larochelle, H. et al.) 4697–4708 (Curran Associates Inc., 2020).
24. Sapsis, T. P. Output-weighted optimal sampling for Bayesian regression and rare event statistics using few samples. *Proc. R. Soc. A* **476**, 20190834 (2020).
25. Mohamad, M. A. & Sapsis, T. P. Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems. *Proc. Natl Acad. Sci. USA* **115**, 11138–11143 (2018).
26. Sapsis, T. P. & Blanchard, A. Optimal criteria and their asymptotic form for data selection in data-driven reduced-order modelling with Gaussian process regression. *Philos. Trans. R. Soc. A* **380**, 20210197 (2022).
27. Kermack, W. O. & McKendrick, A. G. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A* **115**, 700–721 (1927).
28. Anderson, R. M. & May, R. M. Population biology of infectious diseases: Part I. *Nature* **280**, 361–367 (1979).
29. Majda, A. J., McLaughlin, D. W. & Tabak, E. G. A one-dimensional model for dispersive wave turbulence. *J. Nonlinear Sci.* **7**, 9–44 (1997).
30. Cai, D., Majda, A. J., McLaughlin, D. W. & Tabak, E. G. Spectral bifurcations in dispersive wave turbulence. *Proc. Natl Acad. Sci. USA* **96**, 14216–14221 (1999).
31. Zakharov, V. E., Guyenne, P., Pushkarev, A. N. & Dias, F. Wave turbulence in one-dimensional models. *Phys. D: Nonlinear Phenom.* **152-153**, 573–619 (2001).
32. Zakharov, V. E., Dias, F. & Pushkarev, A. One-dimensional wave turbulence. *Phys. Rep.* **398**, 1–65 (2004).
33. Pushkarev, A. & Zakharov, V. E. Quasibreathers in the MMT model. *Phys. D: Nonlinear Phenom.* **248**, 55–61 (2013).
34. Cousins, W. & Sapsis, T. P. Quantification and prediction of extreme events in a one-dimensional nonlinear dispersive wave model. *Phys. D: Nonlinear Phenom.* **280**, 48–58 (2014).
35. Chapelle, O. & Li, L. An empirical evaluation of Thompson sampling. In *Proc. 24th International Conference on Neural Information Processing Systems* (eds Shawe-Taylor, J. et al.) 2249–2257 (Curran Associates Inc., 2011).
36. Nakkiran, P. et al. Deep double descent: where bigger models and more data hurt. *J. Stat. Mech.* **2021**, 124003 (2021).
37. Pickering, E. & Sapsis, T. P. Information FOMO: the unhealthy fear of missing out on information. A method for removing misleading data for healthier models. Preprint at https://arxiv.org/abs/2208.13080 (2022).
38. Sapsis, T., Pipiras, V., Weems, K. & Belenky, V. On extreme value properties of vertical bending moment. In *Proc. 33rd Symposium on Naval Hydrodynamics Osaka, Japan (Virtual)* (2020).
39. Sapsis, T. P., Belenky, V., Weems, K. & Pipiras, V. Extreme properties of impact-induced vertical bending moments. In *Proc. 1st International Conference on the Stability and Safety of Ships and Ocean Vehicles* (2021).
40. Belenky, V., Weems, K., Sapsis, T. P. & Pipiras, V. Influence of deck submergence events on extreme properties of wave-induced vertical bending moment. In *Proc. 1st International Conference on the Stability and Safety of Ships and Ocean Vehicles* (2021).
41. Serebrinsky, S. & Ortiz, M. A hysteretic cohesive-law model of fatigue-crack nucleation. *Scripta Mater.* **53**, 1193–1196 (2005).
42. Khan, R. A. & Ahmad, S. Dynamic response and fatigue reliability analysis of marine riser under random loads. In *Proc. Petroleum Technology Symposium of International Conference on Offshore Mechanics and Arctic Engineering* Vol. 2, 183–191 (ASME, 2007).
43. Chasparis, F. et al. Lock-in, transient and chaotic response in riser VIV. In *Proc. International Conference on Offshore Mechanics and Arctic Engineering*, Vol. 5, 479–485 (ASME, 2009).
44. Lin, W.-M., Zhang, S. & Weems, K. M. Numerical simulations of surface effect ship in waves. In *Proc. 2010 Conference on Grand Challenges in Modeling and Simulation* 414–421 (Society for Modeling and Simulation International, 2010).
45. Li, Z. et al. Fourier neural operator for parametric partial differential equations. In *Proc. International Conference on Learning Representations* (ICLR, 2021); https://openreview.net/forum?id=c8P9NQVtmnO
46. Yang, Y., Blanchard, A., Sapsis, T. P. & Perdikaris, P. Output-weighted sampling for multi-armed bandits with extreme payoffs. *Proc. R. Soc. A* **478**, 20210781 (2022).

47. Hansen, L. K. & Salamon, P. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**, 993–1001 (1990).
48. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proc. 31st International Conference on Neural Information Processing Systems* (eds Guyon, I. et al.) 6405–6416 (Curran Associates Inc., 2017).
49. Gustafsson, F. K., Danelljan, M. & Schon, T. B. Evaluating scalable Bayesian deep learning methods for robust computer vision. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 318–319 (IEEE, 2020).
50. Loshchilov, I. & Hutter, F. SGDR: stochastic gradient descent with warm restarts. In *Proc. International Conference on Learning Representations* (ICLR, 2017); https://openreview.net/forum?id=Skq89Scxx
51. Huang, G. et al. Snapshot ensembles: train 1, get M for free. In *Proc. International Conference on Learning Representations* (ICLR, 2017); https://openreview.net/forum?id=BJYwwY9ll
52. Smith, L. N. No more pesky learning rate guessing games. Preprint at https://arxiv.org/abs/1506.01186 (2015).
53. Gramacy, R. B. & Lee, H. K. H. Adaptive design and analysis of supercomputer experiments. *Technometrics* **51**, 130–145 (2009).
54. Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**, 285–294 (1933).
55. Pickering, E. dnosearch_nature_cs_data (Zenodo, 2022); https://doi.org/10.5281/zenodo.7314144
56. Pickering, E. dnosearch (Zenodo, 2022); https://doi.org/10.5281/zenodo.7312058

Acknowledgements

We acknowledge support from DARPA grant no. HR00112290029, AFOSR MURI grant no. FA9550-21-1-0058 and ONR grants nos. N00014-20-1-2366 and N00014-21-1-2357, awarded to MIT. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank V. Belenky and K. Weems from NSWC at Carderock for support regarding the LAMP code, as well as A. Blanchard for helpful and stimulating discussions around the Bayesian experimental design.

Author contributions

E.P. and T.P.S. conceived the idea and developed the neural operator–BED framework. E.P. implemented the functional search approach for neural operators. G.E.K. contributed the DNO architecture and framework. E.P. conducted the pandemic and rogue-wave simulations and numerical experiments. S.G. conducted ship simulations and numerical experiments. All authors interpreted the results. E.P. wrote the original manuscript. All authors contributed to editing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-022-00376-0.

Correspondence and requests for materials should be addressed to Ethan Pickering or Themistoklis P. Sapsis.

Peer review information *Nature Computational Science* thanks Giovanni Dematteis, Xiang Zhou and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Jie Pan, in collaboration with the *Nature Computational Science* team.