Enabling immersive engagement in energy system models with deep learning

Bruce Bugbee1 | Brian W. Bush2 | Kenny Gruchalla1 | Kristin Potter1 | Nicholas Brunhart-Lupo1 | Venkat Krishnan3

1Computational Sciences Center, National Renewable Energy Laboratory, Golden, Colorado
2Strategic Energy Analysis Center, National Renewable Energy Laboratory, Golden, Colorado
3Power Systems Engineering Center, National Renewable Energy Laboratory, Golden, Colorado

Abstract
Complex ensembles of energy simulation models have become significant components of renewable energy research in recent years. Often the significant computational cost, high-dimensional structure, and other complexities hinder researchers from fully utilizing these data sources for knowledge building. Researchers at National Renewable Energy Laboratory have developed an immersive visualization workflow to dramatically improve user engagement and analysis capability through a combination of low-dimensional structure analysis, deep learning, and custom visualization methods. We present case studies for two energy simulation platforms.

KEYWORDS
high-dimensional data, interactive visualization, neural networks, renewable energy, t-SNE, Tucker decomposition

1 INTRODUCTION

Scientists have access to more complex simulation platforms than ever before. These simulation engines allow researchers to study a wide range of problems ranging from fluid dynamics and power-flow optimization to techno-econometric forecasting. These systems often require significant high performance computing (HPC) resources to run. The computational cost of individual runs combined with the complexity of both input and output data from these simulation engines often inhibit researchers’ ability to fully leverage these models for building domain knowledge. This “pain point” provides a fertile area for research on how to best enable the exploration, assimilation, and communication of scientific data arising from complex simulation environments.

The work presented in this paper reflects efforts undertaken by researchers at the National Renewable Energy Laboratory (NREL) to create and facilitate immersive engagement between domain researchers and their data. The remainder of this paper is organized based on the various aspects we have identified as important factors for improving researcher engagement and efficiency. Section 2 provides background information on two flagship NREL energy simulation engines as well as our work and capabilities surrounding interactive visualization. Section 3 describes work done on gaining meaningful insight into low-dimensional structures from the high-dimensional input and output data associated with these simulation engines. Section 4 details initial efforts at emulating the input-output regression relationship through deep learning. Section 5 highlights how we leverage interactive visualization with emulated simulation models to allow researchers a fully immersive experience. Finally, Section 6 will describe avenues for future work.

2 BACKGROUND

Modern renewable energy research increasingly incorporates complex energy-system-simulation models to study ques-
FIGURE 1 Flowchart detailing the use of reduced-form models (bottom row) and interactive visualization (bottom three rows) to explore energy-simulation models. The top three rows show successive automation of the analysis process: the second row emphasizing agile visualization, while third additionally emphasizes agile construction of input scenarios.

We term this interactive workflow *ensemble steering*. The basic goal of this workflow is to provide a cohesive computational and visualization system that allows domain researchers to directly explore and interact with data from their energy simulations. This includes the development of visualization and statistical methods to highlight hidden structure and grouping characteristics, tools to monitor the impact of various tunable input parameters, and the ability to directly interrogate energy simulations by launching new simulation runs for areas of interest. Oftentimes this last case is difficult given the sheer computational cost associated with a single run of these energy simulation platforms. To be truly immersive, the lag between request and delivery of the new simulation output needs to be less than the order of seconds. In cases where this is infeasible for a given simulation model, a reduced-form statistical model trained to emulate the input-output relationship of the simulation platform can be used to approximate results. Furthermore, additional results from the simulation platform to get the appropriate exact output can be initialized and later integrated to the full ensemble when complete.

2.1 Energy system models

2.1.1 Biomass scenario models

Energy analysts and stakeholders at NREL actively use in-house tools developed for the visualization of generic datasets of multidimensional time series to explore results of biomass supply-chain models such as the Biomass Scenario Learning Model (BSLM) [37], the Biomass Scenario...
Model (BSM) [26], and a waste-to-energy system simulation (WESyS). This suite of simulations uses system-dynamics methodology [35] to model dynamic interactions within the biomass-to-bioenergy supply chain: the models track the deployment of bioenergy given technological developments, government policy, and the reaction of the investment community to those technologies and policies [26]. These are deterministic simulations.

In the work presented in this paper, we explored data from a BSLM sensitivity analysis. The input variables of interest are relative investment levels, government policy, and rate of industrial learning (“learning by doing”) for three competing biofuel technologies (generically labeled A, B, and C, but representing real-life techno-economic pathways) at the pilot-plant, demonstration, and commercial stages of biorefinery development. The outputs are forecast biofuel production levels for each technological pathway (A, B, and C) for each year from 2017 to 2031.

Regional energy deployment system

The Regional Energy Deployment System (ReEDS) model is an electricity system capacity expansion model that develops scenarios of future investment and operation of generation and transmission capacity to meet US electricity demand [9]. ReEDS represents the continental United States with a spatial resolution of 134 balancing areas (where electricity demand and generation are balanced), 356 wind and solar regions (characterizing performance, cost and resource data), and 17 annual time slices (for reduced-form generation dispatch) [18]. ReEDS performs a deterministic, system-wide cost optimization in 2-year solution periods from 2010 out to 2050. Technologies include both conventional (coal, nuclear, gas and oil) and renewable (wind, solar, bio-power, geothermal, hydro) energy production. While the ReEDS model optimizes the technology investments from a system perspective, it also incorporates decisions for the distributed energy resources, such as rooftop solar technologies, through its integration with dGen models [33]. The dGen (aka dSolar) model estimates the yearly anticipated construction of rooftop-solar panel based on customer cash-flow analysis and market-adoption theory, subject to the variability in the market signals and housing capability [8]. The model also considers technology, resource, and policy constraints including state renewable portfolio standards. The outputs from planning model include cumulative capacity of fossil, nuclear, renewable, and storage resource deployments, the transmission expansions, the generator dispatch and fuel needed to satisfy regional demand requirements and to maintain grid system adequacy, total system cost, electricity prices, emissions, and water consumption. Notable studies that have made significant use of ReEDS include the Wind Vision Study [39], a prospective analysis of state Renewable Portfolio Standards impacts [21], and the National Plug-in Electric Vehicles impacts assessment study [24].

In the work presented here, ReEDS model simulations were performed for 18,515 scenarios using NREL’s HPC system. Scenarios embody an experimental design of Latin Hypercube Sampling that varied the cost of renewable energy technologies and resources (geothermal, hydroelectric, rooftop solar, utility solar, and wind), natural gas prices, and electricity demand. This extensive scenario generation facilitated the deep-learning based immersive engagement study reported in this paper. The outputs of interest for this study from all the scenarios are the projected capacity and generation for a variety of technologies estimated at 2 year intervals from years 2016 to 2050.

Interactive visualization

The visualization systems developed for this work have two primary focuses: the development and facilitation of the emulated simulations, and the use of emulated simulations to facilitate rapid rendering of the simulation-output space: one visualization directive helps us understand if our prediction scheme is working, how well it is working, and where it fails, while the other directive uses the prediction tool to steer an ensemble analysis of the simulation. While both tools employ similar visualization techniques, such as parallel coordinates [14], their use of the technologies differs based on the goals of each of the tools. The development of all of the tools draws on previous visualization research in interactive visualization of high-dimensional and multivariate data, ensemble visualization, and ensemble steering.

The interactive visualization of data provides user interfaces to a visualization tool, in order to give the user the ability to explore the data and eventually develop their own view of the data, reflecting knowledge discovery from that exploration. Common interactive techniques include focusing and linking [5] in which the user can focus in on specific data elements or regions, or link data elements in order to reveal commonalities between them. More complex interactions, including brushing [22], in which users broadly select or “brush” regions of interest and this selection may then be reflected in multiple views of the data. As the dimensionality or number of variables within a dataset grows, these interactions become necessary since the data to be visualized outpaces the three spatial dimensions (plus one for the time dimension) that we have to visually display information. The aforementioned techniques are thus typically used to reduce the number of dimensions or variables to display; however, there are a number of alternative visualization techniques that try to alleviate the complexities of high-dimensional data display. For a full treatment of such methods, Buja, Cook, and Swayne have developed taxonomy of such methods that classifies methods based on data analytic task and
interactive view manipulations [4]. Methods directly related to the work presented here are interactive parallel coordinates [36] and the use of t-Distributed Stochastic Neighbor Embedding (t-SNE) [20], both of which are described in more detail in Section 3.

While the techniques mentioned above deal with the interactive exploration of high-dimensional data, more relevant techniques deal with the interrogation and visualization of ensemble datasets. Ensembles specifically refer to multiple runs of a simulation using different sets of perturbed input parameters and initial conditions and are used to mitigate the effects of uncertainty and span the space of outcome results. These types of datasets are becoming more and more popular and their inherent complexity leads to many challenges in visualizing them [25]. Their multifaceted nature refers to the heterogenous data characteristics; they are often multivariate and spatiotemporal, multimodal—meaning they may combine multiple data sources or simulation models, and may have discrepancies in dimensionality, grid size, or structure, all of this leading to non-trivial issues in handling this type of data. A survey of the challenges and techniques for multifaceted data can be found in the survey by Kehrer and Hauser [15]. Systems designed to handle ensemble data, specifically originating from weather and climate simulations include Ensemble-Vis [29] and Noodles [31], both of which use multiple windows linked via user interactions and rely heavily on 2D charts to express differences between ensemble members. Volume rendering and isosurfacing are also very popular techniques [28,27], but they are often used to express uncertainty within datasets that also have an inherent spatial domain. Also closely related is the exploration of parameter-spaces [32], where a visualization tool is developed exclusively to better understand how changes in the input parameters relate to changes in the output space. Such systems have led to the steering of simulations in which the user can interactively choose parameter settings that are then used to launch new simulations runs [38,30,23]. Our work here resembles these steering methods in that new parameter settings initiate new simulations, except our method first estimates the simulation space giving the user a general idea of the relationships between parameter settings and output results. From this prediction, the user can then choose regions of interest from which to spawn simulations, adding to the ensemble dataset.

Our interactive visualization work utilizes a variety of visualization mediums from standard desktops to state-of-the-art visualization systems, including high-resolution displays and an immersive virtual environment. High-resolution displays, such as NREL’s large-scale 14-megapixel display wall, provide the visual real estate to layout and view multiple variables from multiple ensembles simultaneously. Immersive virtual environments allow users to explore and interact with their data physically in three-dimensional space. The immersive virtual environment at NREL provides a tracked stereoscopic environment composed of six projectors that illuminate two surfaces—a wall and a floor—creating a 10-megapixel $5 \times 2.5 \times 2$ m immersive volume. The extra degrees of freedom provided by immersive environments allow us to present high-dimensional data in new ways and provide visibility into complex multivariate relationships.

### 3 | EXPLORING ENERGY SYSTEMS MODELS

Exploration and discovery challenges refer to the set of problems of effectively finding what portions of data space for the target simulation are of most interest. This can range from simply identifying important, known dimensions through domain expertise to uncovering hidden, low-dimensional latent structure using more advanced analysis. Analysts can often be overwhelmed by the sheer scope of data available. Aiding in determining important dimensions for them to focus on during initial investigations is important in overcoming this “drowning in data”.

Energy system models such as ReEDS and BSM have dozens or hundreds of input and output variables that interest energy analysts. Even the simplest national-level-aggregate analyses using ReEDS, for example, involve varying a half dozen input parameters (eg, technology costs, energy prices, and energy demand) and examining two dozen output metrics (eg, energy production capacity and production from numerous technologies) over biannual time series (eg, 2016-2050), resulting in a data space with more than 500 dimensions (ie, the number of input variables plus the product of the number of time values and the number of output variables). Because these models represent highly designed systems with many physical, temporal, economic, resource, and behavioral constraints, their range (actual values from the model) is much smaller than their codomain (theoretically possible values). In ReEDS, for example, the production capacity for a technology like wind power generally increases monotonically; in BSM, the sum over technological pathways of the cellulosic biofuel production is constrained by the land available on which cellulosic crops might be grown. Thus, the input-output datasets for these models form lower dimensional manifolds embedded in higher dimensional spaces. The existence of such low-dimensional structure improves opportunities for visually exploring model results and the characteristics of the structures themselves provide insights into the model results.

In the remainder of this section we highlight one technique, combining tensor-decomposition with embedding in three dimensions, that we have found particularly helpful in exploring high-dimensional output from energy system models. This does not, of course, preclude the use of standard exploratory data analysis techniques (correlation analysis, scatterplot matrices, variance decomposition, principle
components analysis, nonlinear dimensionality reduction, topological data analysis, etc.).

One major challenge for exploring energy system models is identifying, constructing, and understanding the low-dimensional embeddings of model results. Exploring the features and structure of the high-dimensional time series output from energy system models poses a cluster of related challenges: (a) reducing the dimensionality, (b) visualizing the embedded submanifolds or features, and (c) respecting the temporal nature of the output. Regarding temporality, we would like to avoid methods, like principal components analysis (PCA), which would mix the temporal and metric dimensions as their reduce dimensionality. (We do not include input dimensions/variables when identifying features because the structure of the input space is determined by whatever sampling or design method was chosen for generating the ensemble of the energy system model simulations: ie, its structure reflects the analyst’s choice rather than intrinsic properties of the model.)

The t-SNE algorithm is a general and effective method for visualizing moderately high dimensional data, but it is not particularly well suited to dealing with datasets having hundreds of dimensions and it does not treat time series especially [20]. Furthermore, selecting two dimensions as the target for visualizations based on t-SNE overly constrains the algorithm for energy system models with particularly rich or complicated sets of features/submanifolds. Selecting three dimensions as the target for t-SNE substantially remedies the latter shortcoming because having a third dimension allows the algorithm more flexibility in the situation where the model’s features locally relate to one another in a “non-planar” manner.

In order to handle the temporal aspect of these models and to preliminarily reduce the dimensionality of the data to a size suitable for the application of t-SNE, we apply a Tucker decomposition [16] to the dataset prior to applying t-SNE. We choose to use a Tucker decomposition here, but any method in the family of tensor-based, multilinear principal components analysis (MPCA) might be substituted. Prior to applying the Tucker decomposition, we scale the data using a min-max scaler. The tensor used as input to the Tucker decomposition is indexed by ensemble member (1, … , n, where n is the size of the ensemble of results), output variable (1, … , m, where m is the number of output variables), and time slice (1, … , t, where t is the number of times reported). We decompose that tensor \( X \in \mathbb{R}^{n \times m \times t} \) into a core tensor \( Y \in \mathbb{R}^{n \times m' \times t'} \) and factor matrices \( M \in \mathbb{R}^{t \times m'} \) and \( T \in \mathbb{R}^{t' \times t} \), reducing the metric dimensions from \( m \) to \( m' \) and the temporal dimensions from \( t \) to \( t' \):

\[
X \approx Y \otimes M \otimes T. \tag{1}
\]

Thus the original dimensionality of the \( n \) ensemble members is reduced from \( mt \) to \( m't' \). Because of the orthogonality of factors imposed by the Tucker decomposition, the proportion of the variance explained is \( \text{Var}Y/\text{Var}X \), just as in the case in PCA [2]. Relying upon this variance property, a suitable choice of \( m' \) and \( t' \) allows the subsequent practical application of t-SNE to \( Y \). When output metrics are highly linearly correlated, the approximation holds well for \( m' \ll m \); when temporal behavior is similarly regular, the approximation holds well for \( t' \ll t \).

Although the aforementioned procedure embeds the raw data into a visualization that organizes the features in three dimensions, it does not clearly characterize the properties of the features. To do that, we map the value of an analyst-defined selection of input or output variables to the color, size, or other attributes (glyphs, etc.) of the embedded data. If such input or output variables correspond to a local organizing principle for the embedding, then the visualization will exhibit clear local trends (in color, size, etc.) associated with those variables. Figures 2 and 3 show example of embeddings, though local trends in color and structure are far easier to see in three dimensions than in these two-dimensional projections. Figure 2 shows clear separation between high levels of biofuel production in a particular technological pathway (orange ensemble members) vs low levels (blue). We further present interactive visualizations in Section 5.2.

Anecdotally, analysts have observed intriguing commonalities among visualizations for models relying on similar modeling approaches and for energy systems of similar maturity. Optimization models (eg, linear programs) tend to result in connected shapes with continuously varying trends in variables whereas dynamical models (eg, coupled ordinary differential equations) tend towards disconnected or bifurcating shapes. Models of mature energy systems (eg, the electric power system) tend towards continuity, but models of emerging technologies (eg, biofuels or automated vehicles) tend towards bifurcation.

4 EMULATING ENERGY SYSTEM MODELS

Emulating energy system models is, at its core, a regression problem. For a given model, our primary goal is to accurately map the regression relationship between some set of input parameters and some complex set of output data. The only requirement we place on this process is that evaluation of a trained model must be fast enough to facilitate an interactive exploration workflow. This opens up a large toolbox of regression methods that could be appropriate for a given simulation platform, ranging from simple linear models to advanced deep neural networks. For the case studies involving BSLM and ReEDS, we specifically investigated deep neural networks.
FIGURE 2   Ensemble of $n = 50,000$ Biomass Scenario Learning Model simulations with $m = m' = 3$ metric dimensions, temporal dimensions reduced from $t = 10$ to $t' = 5$, and with ensemble members colored by the production of biofuel using a particular technological pathway. The visual embedding has organized regions where much (colored orange) biofuel is produced by the pathway against a background where little (colored purple) biofuel is produced by that pathway. The diagonal shows as histogram of the density of observations along each of the three embedding axes; the lower left scatterplots show projections of the embedding on to each pair of the three axes; and the upper right contour plot shows projections of the embedding onto each pair of the axes. In three dimensions (see Figure 11), the point cloud has a dense ball-shaped structure with “horns” and a small disconnected ball between the horns. Local trends in color and structure are much easier to see when these data are plotted in three dimensions.

FIGURE 3   Ensemble of $n = 18,515$ Regional Energy Deployment System simulations with metric dimensions reduced from $m = 24$ to $m' = 12$, temporal dimensions reduced from $t = 10$ to $t' = 5$, and with ensemble members colored by the generation of electricity by natural-gas peaking units. The visual embedding has organized regions where little (colored purple) electricity from natural-gas peaking units is generated, against a background (mixed purple and orange) where that local trend does not persist. The diagonal shows as histogram of the density of observations along each of the three embedding axes; the lower left scatterplots show projections of the embedding on to each pair of the three axes; and the upper right contour plot shows projections of the embedding onto each pair of the axes. In three dimensions (see Figure 11), the point cloud is saucer-shaped. Local trends in color and structure are much easier to see when these data are plotted in three dimensions.

Our primary efforts for emulating energy systems models focused on applying variations on multilayer neural network architectures. Recent advances in model construction, GPU computation, and programming frameworks have made it easier than ever to construct, train, and deploy complicated neural network architectures to solve both regression and classification tasks [19]. Since our primary goal is predictive fidelity to the underlying energy simulation, the lack of interpretability of the multiple layers of nodes and their nonlinear connections is not particularly concerning for this case. The TensorFlow backend [1] to Keras [7] was used for performing the computations. Figure 4 shows the Artificial Neural Network architecture used for these emulations. The BSLM and ReEDS models have different numbers of input and output variables, but the each hidden layer consists of 12 neurons with rectified linear units for activation. The output
Figure 4  Artificial neural network architecture for approximations to Biomass Scenario Learning Model and Regional Energy Deployment System models. The network nodes are either simple activation functions, in the case of a plain network, or highway functions that transform the activation and carry information from the previous layer, in the case of a highway network.

Figure 5  Estimated vs Biomass Scenario Learning Model forecasted output for Technology A using the baseline 1-nearest neighbor model (simple lookup table). The displayed data are 1000 sampled runs from an independent testing dataset. BSM, Biomass Scenario Model

layer uses linear activation and the mean squared error is used for the loss function. All layers are fully connected. Fully connected plain networks are used for ReEDS. In the highway networks [34] that we use for the BSLM, the hidden layers route information not only through the activation function (via the “transform” behavior), as in a plain neural network, but also bypass the activation function (via the “carry” behavior); the default Keras transform and carry behavior is used.

4.1  Biomass Scenario Learning Model

We experimented with multiple architectures for emulating the BSLM energy simulation. As mentioned in Section 2.1.1, the BSLM model maps nine continuous valued inputs representing technological learning ratios at different developmental stages to a longitudinal forecast for biofuel output for each technology. Rather than trying to build a single model to entirely capture this output space, our initial efforts instead partition the output data and train separate models for each combination of output time slice and technology. While training separate models potentially ignores interesting correlation structure in the output space, the ability to train these models in parallel made them ideal for initial prototyping work.

For BSLM we considered two classes of architectures: standard fully connected networks and highway networks [11, 34]. Highway networks have the advantage of included gated highway connections that allows for persistent state information from previous layers to flow unimpeded across multiple layers. This adaptation, akin to recent advances in residual networks (ResNets) and similar to the gating structure of long short term memory networks allow for the training of very deep fully connected networks where traditional
methods would fail [12,13]. We omit detailed discussion of architecture and hyperparameter tuning for brevity. The main takeaway for our emulation efforts on BSLM can be seen in Figures 5 and 6. When compared to a simple nearest neighbor lookup table in Figure 5, the ensemble of our neural network architectures in Figure 6 demonstrates noticeable improvements in accuracy on an independent test set. Early years (2020 and 2022) demonstrate difficulty in fitting for all methods. This is primarily due to the biofuel output not "taking off" yet, given the particular scenario of investment ratios.

As always there are general concerns of overfitting the data but multiple comparisons to held out BSLM test sets indicate that these networks generalize well enough to support our immersive engagement efforts. This ANN-based emulator runs three orders of magnitude faster than the full simulation (tens of milliseconds instead of tens of seconds).

4.2 | Regional Energy Deployment System

Our efforts emulating the ReEDS energy simulation follows a similar course as described in Section 4.1. Here we partition our output space into a separate model for each longitudinal forecast associated with a unique technology and output type combination. For example, a single model is built to
map our various input ratios concerning technology scenarios (described in Section 2.1.2) for the entire forecast of residential solar capacity while a separate model is trained for residential solar generation forecasts. We limit our initial exploration to simple fully connected networks for these models, which proved to be quite accurate in these cases even with the higher dimensional output space associated with the full longitudinal forecast. Figures 7 and 8 highlight the relative accuracy of our emulators for a select set of ReEDS capacity forecasts. This ANN-based emulator runs four orders of magnitude faster than the full simulation (fractions of a second instead of hours).

5 | INTERACTING WITH ENERGY SYSTEM MODELS

5.1 | Comparison of real and predicted BSLM simulation results

To verify the results of our model predictions, we have developed both desktop and immersive visualization tools that display and compare simulation results. As shown in Figure 9, our desktop tool displays both simulated and real results along with filters to select or highlight regions of interest. The tool has multiple methods for displaying the data: in the figure a parallel coordinates plot shows each of the output parameters, including year, real, and estimated results; additionally, a table of data values can be displayed to allow for direct inspection. Both views have a user interface that allows the user to filter by model type and colormap or to exclude data by selecting or unselecting characteristics including model type, name, data ID, or parameter. The parallel coordinate allows the user to select ranges of years and data ranges via moussing over values to create rectangular selection regions. The table view allows users to sort any of the table columns. Finally, the transparency and sampling rate in the user interface allows the user to refine the visual presentation of the display, a feature most effective on large datasets to help reduce visual clutter.

Immersive visualization provides an alternative view for exploring the fits of the model predictions (Figure 10). Users interactively choose three dimensions of interest from the simulation through a web interface. We then plot the real results as temporal trajectories in our immersive space, showing how each run evolved through time. We represent the error of the predicted results as an ellipsoid placed on each discrete time point along the trajectory, scaled by the relative error of the predicted result in the three cardinal dimensions. The immersive visualization provides a natural interface to explore where in the parameter space our prediction scheme is working and where it fails.

5.2 | Interactive, immersive exploration of Tucker-decomposed t-SNE visualizations

Figure 11 shows an interactive web-browser visualizer linked (Figure 12) to 3D immersive visualization of the Tucker-decomposed t-SNE visualizations described in Section 3. Users can select a dataset, the number of metric dimensions $m'$, the number of temporal dimensions $t'$, the t-SNE perplexity, and the variable for coloring the ensemble members. As these selections are made, the browser-based projections shown in Figure 11 and the immersive 3D visualization shown in Figure 12 update to reflect the choices. Keyboard controls allow users to rapidly cycle through each of the settings, making it practical to view dozens of visual presentations in a short time. Analysts typically vary the choice of $m'$, $t'$, t-SNE parameters such as perplexity, and colored variable until clear local patterns and trends become...
FIGURE 9  Web-browser-based visualization tool for comparing predicted and real simulation results to aid in verifying our approach. The parallel coordinates view shows relationships between output results by plotting the data value on each of the dimensional axes. A tabular view (not shown here) directly displays data results. Both display types can be filtered via the tools on the left-hand of each image.

FIGURE 10  Photograph of the immersive comparison of Regional Energy Deployment System (ReEDS) results. Each line represents a real temporal trajectory through three dimensions of the ReEDS parameter space. An ellipsoid represents the error of the prediction for each point in time apparent; then they will often cycle through the colored variables as they identify narratives describing the local feature and the global trends.

The decomposition and t-SNE software is written in Python and the visualizer in R. Both will be released open-source. We use Tensorly [17] for the Tucker decomposition and scikit-learn [17] for t-SNE.

5.3 Interactive, immersive approximate ensemble simulations

We have also connected the approximated ReEDS model to our immersive 3D “parallel planes” visualization [3,6]. Similar to parallel-coordinates plots in 2D, the 3D parallel planes encode input and output variables, and polylines connect the coordinate values for each ensemble member (Figure 13). Users can brush regions of input or output planes to highlight and explore existing data. Users can also brush regions of input planes to specify a multidimensional sampling of ensembles with the brushed input variables. We use the approximate model to generate predictive results of these new ensembles. The new predictions appear almost immediately once the user has requested their creation, as compared to the 6-hour delay that would be required for a full ReEDS simulation to complete. This visualization has been used successfully with configurations from two planes.
FIGURE 11  Screenshot of a web-browser-based control and visualization of the Tucker-decomposed t-Distributed Stochastic Neighbor Embedding algorithm described in Section 3. Users can select datasets and algorithmic parameters using keyboard or mouse, resulting in the revised visualization being displaced on the browser and in the immersive 3D space shown in Figure 12.

FIGURE 12  Photograph of the 3D immersive visualizations of the Tucker-decomposed t-Distributed Stochastic Neighbor Embedding algorithm described in Section 3 and controlled via a web browser (Figure 11): the left side corresponds to the 2D visualization in Figure 2 and the right side to that in Figure 3.

(ie, two input variables and two output variables) to 10 planes (ie, seven input variables and 12 output variables). The main strengths of the parallel-planes technique are for exploring high-dimensional correlations among input and output variables, for generating and testing hypothesis about model/system behavior, for model debugging, and for planning full-model simulation studies.

6 | CONCLUSION

The primary goal of statistics and data science in a scientific laboratory is to provide domain researchers with the tools needed for gaining insight from data and increasing knowledge. As data sources become more complex and advanced simulation platforms become increasingly important for researchers, policy architects, and others, methods that allow stakeholders to immerse themselves and interact with their data are increasingly valuable. The work we present here focused on three key components of an immersive workflow: exploration, emulation, and visualization. By combining ideas from low dimensional structure embedding, deep learning, and advanced visualization techniques, we have been able to prototype advanced analytics platforms for two important energy simulation models. Preliminary engagement with domain researchers has reinforced the usefulness of our work for better understanding input parameters, policy scenario development, designing next generation simulation platforms, and engagement with upper management stakeholders. Future work will aim to extend all three areas of focus discussed here as well as expand our work to support energy simulations from other domains across the renewable energy space.

Our future work on the Tucker-decomposed t-SNE exploratory visualizations emphasizes the addition of linked views to nonembedded projections of the data, more sophisticated brushing of the data, animated colormaps, tagging of features with user-supplied annotations, and coordinating views on multiple display devices for simultaneous, geographically separated users. The ANN-based emulations are being further developed as interactive tools and packaged as efficient substitutes for the full simulations.

A critical component of driving immersive engagement through regression models that emulate energy simulations is having an accurate understanding of how well these emulators perform. Apart from being good statistical practice, providing uncertainty estimates for predicted output has shown to be important for getting domain scientists to buy in to the usefulness of the these tools. This is somewhat at odds with traditional neural network use cases where prediction is the primary goal and the stacked nature of their layers and non-linearities make interpretability and uncertainty measurement
Increasingly difficult. Recent work from Gal and Ghahramani explores the idea of using the results of stochastic feed forward passes through multiple dropout layers of a trained neural network to represent the model uncertainty via an approximation of Gaussian processes [10]. We plan to adapt the implementation of this method described in ref. [40] to provide rough estimates of model uncertainty associated with our emulated models.

ACKNOWLEDGMENTS

This work was authored by the National Renewable Energy Laboratory (NREL), operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. This work was supported by the Laboratory Directed Research and Development (LDRD) Program at NREL. The views expressed in the article do not necessarily represent the views of the DOE or the US Government. The US Government retains and the publisher, by accepting the article for publication, acknowledges that the US Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for US Government purposes. A portion of this research was performed using computational resources sponsored by the Department of Energy's Office of Energy Efficiency and Renewable Energy and located at the National Renewable Energy Laboratory. The authors acknowledge Steven Smith of Los Alamos Visualization Associates (LAVA) for helpful conversations on strategies for interactively visualizing high-dimensional datasets.

REFERENCES

1. M. Abadi et al., Tensorflow: A system for large-scale machine learning, OSDI 16 (2016), 265–283.
2. G. I. Allen, Regularized tensor factorizations and higher-order principal components analysis, arXiv:1202.2476, 2012.
3. N. Brunhart-Lupo, B. W. Bush, K. Gruchalla, and S. Smith, Simulation exploration through immersive parallel planes, 2016 Workshop on Immersive Analytics (IA), IEEE, 2016, pp. 19–24.
4. A. Buja, D. Cook, and D. F. Swayne, Interactive high-dimensional data visualization, J. Comput. Graph. Stat. 5(1) (1996), 78–99.
5. A. Buja, J. A. McDonald, J. Michalak, and W. Stuetzle, Interactive data visualization using focusing and linking, Visualization, 1991. Proceedings of IEEE Conference on Visualization'91, IEEE, 1991, pp. 156–163.
6. B. Bush, N. Brunhart-Lupo, B. Bugbee, V. Krishnan, K. Potter, and K. Gruchalla, Coupling visualization, simulation, and deep learning for ensemble steering of complex energy models, 2017 IEEE Workshop on Data Systems for Interactive Analysis (DSIA), IEEE, 2017, pp. 1–5.
7. F. Chollet et al., Keras (2015). https://keras.io
8. W. Cole et al., Interactions of rooftop PV deployment with the capacity expansion of the bulk power system, Appl. Energy 168 (2016), 473–481.
9. K. Eurek et al., Regional energy deployment system (ReEDS) model documentation: Version 2016, National Renewable Energy Laboratory (NREL), Golden, CO, 2016.
10. Y. Gal and Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, International Conference on Machine Learning, 2016, pp. 1050–1059.

ORCID

Bruce Bugbee https://orcid.org/0000-0002-5396-1791
11. I. Goodfellow et al., Deep learning, Vol 1, MIT Press, Cambridge, 2016.
12. K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
13. S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural Comput. 9(8) (1997), 1735–1780.
14. A. Inselberg and B. Dimsdale, Parallel coordinates for visualizing multi-dimensional geometry, in Computer graphics, Springer, Tokyo, 1987, 1987, 25–44.
15. J. Kehrer and H. Hauser, Visualization and visual analysis of multifaceted scientific data: A survey, IEEE Trans. Vis. Comput. Graph. 19(3) (2013), 495–513.
16. T. G. Kolda and B. W. Bader, Tensor decompositions and applications, SIAM Rev. 51(3) (2009), 455–500.
17. J. Kossaifi, Y. Panagakis, and M. Pantic, Tensorly: Tensor learning in python, arXiv:1610.09555, 2016.
18. V. Krishnan and W. Cole, Evaluating the value of high spatial resolution in national capacity expansion models using ReEDS, Power and Energy Society General Meeting (PESGM), 2016, IEEE, 2016, pp. 1–5.
19. Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature 521(7553) (2015), 436–444.
20. L. V. D. Maaten and G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008), 2579–2605.
21. T. Mai et al., A prospective analysis of the costs, benefits, and impacts of us renewable portfolio standards, Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA, 2017.
22. A. R. Martin and M. O. Ward, High dimensional brushing for interactive exploration of multivariate data, Proceedings of the 6th Conference on Visualization’95, IEEE Computer Society, 1995, p. 271.
23. K. Matković et al., Visual analytics for complex engineering systems: Hybrid visual steering of simulation ensembles, IEEE Trans. Vis. Comput. Graph. 20(12) (2014), 1803–1812.
24. M. Melaina et al., National economic value assessment of plug-in electric vehicles: Volume I, NREL (National Renewable Energy Laboratory (NREL), Golden, CO, 2016.
25. H. Obermaier and K. I. Joy, Future challenges for ensemble visualization, IEEE Comput. Graph. Appl. 34(3) (2014), 8–11.
26. S. Peterson et al., Overview of the biomass scenario model, National Renewable Energy Laboratory (NREL), Golden, CO, 2015.
27. T. Pfaffmoser, M. Reitinger, and R. Westermann, Visualizing the positional and geometrical variability of isosurfaces in uncertain scalar fields, Comput. Graph. Forum 30(3) (2011), 951–960. https://doi.org/10.1111/j.1467-8659.2011.01944.x.
28. K. Pöthkow and H.-C. Heg, Positional uncertainty of isosurfaces: Condition analysis and probabilistic measures, IEEE Trans. Vis. Comput. Graph. PP(99) (2010), 1–15.