

Human-Machine Collaboration for the Visual Exploration and Analysis of High-Dimensional Spatial Simulation Ensembles

Mai Dahshan¹, Nicholas F. Polys², Leanna House³, Karim Youssef² and Ryan Pollyea⁴

¹*School of Computing, University of North Florida, U.S.A.*

²*Department of Computer Science, Virginia Tech, U.S.A.*

³*Department of Statistics, Virginia Tech, U.S.A*

⁴*Department of Geosciences, Virginia Tech, U.S.A*

Keywords: Simulation Ensembles, Spatial Data, Visual Analytics, Large Scale Visualization, Gaussian Process.


Abstract: Continuous improvements in supercomputing have given scientists from various fields the ability to conduct large-scale multi-dimensional numerical simulation ensembles. A simulation ensemble involves running multiple simulations, each with slight variations in model settings, such as input parameters, initial conditions, or boundary values. Exploring and analyzing these ensembles facilitates understanding parameter sensitivity and the correlations between different ensemble members. To capture these relationships, visual analytical tools are used to extract important features from the ensemble. In many cases, however, these visualizations highlight the differences in the ensemble using aggregated or descriptive statistics, ignoring the correlations and local differences between different spatial regions, which could hinder the exploration process. This paper proposes a visual analytical approach, SpatialGLEE, to interactively explore the spatial variability in the simulation ensemble. The proposed approach uses Gaussian Process Regression (GPR) and Semantic Interaction (SI) to help scientists explore the impact of input parameters on the ensemble and find the commonalities and differences across ensemble members and regions of interest (ROI). GPR models the spatial correlation structure in the ensemble. The modeled data is then inputted into the visualization pipeline for analysis and exploration with SI. The effectiveness of SpatialGLEE is demonstrated using a real-life case study.


1 INTRODUCTION


Numerical simulations are used in many scientific domains such as geosciences, meteorology, or computational fluid dynamics (Winsberg, 2013; Cappello et al., 2015). These simulations help in understanding complex real-world phenomena. However, determining the optimal initial conditions and model inputs is difficult due to the complexity of the studied phenomena. To address this uncertainty, multiple runs are carried out by slightly modifying initial conditions or model parameters, resulting in an ensemble (Wang et al., 2016; Potter et al., 2009). An ensemble allows scientists to investigate commonalities and differences across runs, determine parameter sensitivity, and find optimal settings. Continuous improvements in computing power allow performing large-


scale simulation ensembles on high-resolution grids in a few hours. However, the analysis and exploration of large ensembles still pose challenges. Therefore, an appropriate representation of ensembles is needed for a more intuitive understanding of the simulated model.


Visualization has a crucial role in understanding large volumes of data at a glance. Visual exploration and analysis of multidimensional ensembles enable scientists to gain insight, identify hidden patterns, and make discoveries, contrasting with traditional manual analysis methods that are exhausting and error-prone. The majority of ensemble visualization literature focuses on analyzing aggregated or sampled ensemble members (Wang et al., 2018; Athawale et al., 2020; Chen et al., 2019). While these visualization approaches have shown promising results, their high abstraction involves losing much ensemble data, which could potentially hide important patterns or trends. Moreover, ensemble members and their parameters cannot be directly examined. Additionally, many of these approaches do not account for spatial characteristics in the data and assume prior knowledge of data

^a  <https://orcid.org/0000-0002-5758-4890>

^b  <https://orcid.org/0000-0002-8503-970X>

^c  <https://orcid.org/0009-0003-2848-4131>

^d  <https://orcid.org/0000-0003-4544-9613>

^e  <https://orcid.org/0000-0001-5560-8601>

patterns, restricting the analysis process.

By collaborating closely with domain experts, we observed their interest in understanding spatial variation in the ensemble, identifying and comparing major patterns between ensemble members, and tracking parameter sensitivity and optimization. Exploring and analyzing spatial variability involves identifying spatial regions that exhibit variability and pinpointing common features and patterns within these regions. Therefore, we propose an interactive approach, SpatialGLEE, for simultaneously exploring multidimensional spatial ensemble members and their parameters. Our approach helps scientists understand the ensemble by exploring "what-if" scenarios to validate hypotheses about the ensemble and its parameters. It involves examining ensemble variability, selecting subsets of ensemble members, and selecting spatial sub-regions within ensemble members for further analysis. To achieve this, we use Gaussian Process Regression (GPR) to encode the spatial structure of each ensemble member, preserving the spatial trends, outliers, variability, and autocorrelation in the data. This paper focuses on the exploration and analysis of 2D spatial ensembles.

This paper presents SpatialGLEE, an expanded interactive approach built on the GLEE visualization tool (Dahshan et al., 2020) to explore and analyze multidimensional spatial ensembles. SpatialGLEE specifically addresses spatial variability, unlike GLEE, which focuses on derived or summary statistics. SpatialGLEE has two main steps: 1) modeling input parameters and simulation outputs while preserving the spatial structure; 2) interactive analysis and exploration of ensemble members, ROI, and simulation inputs and outputs. SpatialGLEE leverages statistical modeling and visual analytics techniques into interactive coordinated visualizations to help scientists simultaneously explore and make sense of spatial ensemble and parameter spaces, considering scientists' visual reasoning, complexity, and structure of data. This enables scientists to visually determine complex insights, including spatial correlations among ensemble members and ROI, as well as parameter sensitivity and optimization. To summarize, the following are the contributions of the paper:

- Refining GLEE's SI pipeline and developing a visualization pipeline to support SpatialGLEE in exploring spatial ensemble and parameter spaces.
- Demonstrating that GPR can preserve spatial characteristics in spatial ensembles, leading to potentially meaningful scientific insights.
- Implementing a parallelized version of maximum likelihood estimation (MLE) for GPR to enhance

scalability with spatial grid sizes, achieving a $21\times$ speedup.

- Demonstrating the effectiveness of our proposed approach in exploring and analyzing multidimensional spatial ensembles using real-world data and domain expert feedback.

2 RELATED WORK

Ensemble visualization approaches have been proposed to examine ensemble member correlations, parameter optimization, and sensitivity (Wang et al., 2018). Common methodologies for ensemble visualization either aggregate ensemble members by calculating the statistical properties of the ensemble (Potter et al., 2009) or transform ensemble members into more abstract representations (i.e., isocontours, iso-surfaces, pathlines, streamlines, etc.) using major trends in the ensemble (Kumpf et al., 2021; Zhang et al., 2020). The former methodology is most relevant to our approach. Aggregated ensembles are usually represented using different statistical displays, including but not limited to box plots (Mirzargar et al., 2014), parallel coordinates (Wang et al., 2016), and line charts (Demir et al., 2014). However, these techniques often hinder many details about the ensemble, resulting in the loss of significant information about the data. Moreover, they are prone to visual cluttering.

To overcome these limitations, improved techniques were developed to reveal variations and detailed information about the data distribution using histograms (Ahmed et al., 2019), statistical dependencies (Li et al., 2017), and circular treemaps (Huang et al., 2023). Moreover, clustering techniques have been employed to detect major patterns by grouping location points that follow certain distributions (Shu et al., 2016). However, these techniques are limited in their application to multidimensional parameter settings. Therefore, they would not provide scientists with a complete picture of both ensemble and parameter spaces.

Recently, ensemble visualization approaches have tried to address the exploration and analysis of multidimensional spatial ensembles. Several efforts attempted to capture ensemble spatiality using diverse techniques, including confidence intervals (Vittinghoff et al., 2022), hyper-slicer (Evers and Linsen, 2022), neural network-latent-based surrogate model (Shi et al., 2022), deep neural networks (Huesmann and Linsen, 2022), critical points (Favelier et al., 2018), function plots, (Fofonov and Linsen, 2018), similarity measure (Fofonov and Linsen,

2019) and uncertainty calculation (Liu et al., 2018). However, the majority of these approaches are primarily focused on parameter space exploration or analyzing a few ensemble members at once, with less emphasis on simultaneously exploring both parameter and ensemble spaces (Orban et al., 2018). Therefore, our approach focused on the exploration of spatial ensembles by integrating visualization with a human-machine collaboration technique to empower the visual analysis of parameter and ensemble spaces.

3 APPROACH

3.1 System Design

SpatialGLEE is designed to help scientists gain insights and find discoveries about simulated data for more effective exploration and analysis. Therefore, our proposed approach and its manifestation in a visual analytics tool result from a long-term collaboration with geoscientists. We studied scientists' conventional analysis workflows through interviews and focus groups while developing the method and tool. This identified the analysis tasks scientists need to understand spatial ensembles. The main analysis goals are as follows:

Goal 1: Parameter Optimization and Sensitivity Analysis. Parameter sensitivity analysis examines the relationship between ensemble members and the model parameters. Scientists aim to explore the parameter space to identify 1) key input parameters that contribute more to explaining simulation outputs and those with little or no impact; 2) the association between multiple input parameters; and 3) input parameter correlation(s) with spatial trends or features. Concurrently, parameter optimization determines the optimal parameter settings for given objectives.

Goal 2: Interactive Exploration and Comparison of Ensemble Members. Exploring individual ensemble members can uncover the commonalities and differences between different groups. Scientists explore the ensemble space to identify and interpret: 1) the locations and reasons behind similarities or differences among ensemble members; 2) prevailing patterns, trends, and anomalies; and 3) the spatial correlation and variability among subsets of ensemble members.

Goal 3: Interactive Exploration of Subsets of Ensemble Members and ROI. Understanding the ensemble's inherent spatial structure enables scientists to investigate the dynamics of simulated

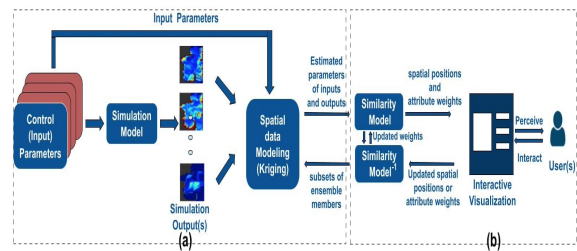


Figure 1: The workflow of our proposed visual analytical approach. Our approach has two main steps: a) Statistical modeling of the ensemble while considering spatial characteristics in the data. b) Interactive visual exploration and analysis of ensemble members, simulation parameters, and spatial patterns.

models. Scientists aim to understand and analyze: 1) ROI's spatial characteristics within the ensemble to find features and patterns across members; 2) how specific parameter or parameter settings affect ROI; and 3) subsets of ensemble members to determine if the relationship between input parameters holds for the entire ensemble or only within specific subsets, and vice versa.

Goal 4: Interactive Exploration and Comparison of Spatial Distributions. Exploring the statistical attributes of raw spatial data to uncover correlations and static properties. This exploration involves analyzing the data distribution to test hypotheses and verify findings.

3.2 System Overview

This section introduces our visual exploration framework for the multidimensional spatial ensemble. The proposed framework and its visualization components align with the aforementioned tasks. Figure 1 provides a high-level overview of our approach. Our approach begins with an ensemble E of M members. Individual ensemble members consist of input parameters and simulation outputs (i.e., grid). Initially, we model the simulation ensemble using GPR to estimate spatial parameters that characterize each member's input parameters and simulation outputs. GPR employs local MLE to determine various spatial process features. These features are spatially smoothed to capture the correlation structure between nearby grid points. GPR does not require prior knowledge of the ensemble member distribution. Thus, it can uncover each ensemble member's variability, main trends, and autocorrelations. The estimated spatial parameters are subsequently inputted into the visualization pipeline of SpatialGLEE.

SpatialGLEE's coordinated multi-views (Figure 2), coupled with their supported interaction tech-

niques, empower scientists to explore and analyze multidimensional spatial ensembles. Ensemble view (Figure 2a) visualizes ensemble members by projecting them from higher-dimensional space to lower-dimensional space (i.e., 2D) via a projection technique (e.g., MDS, t-SNE, etc.), using estimated spatial parameters for simulation outputs and input parameters and weights associated with them. The positioning of ensemble members in the ensemble view reflects relative distances, where members with similar estimated attributes are placed close together, while those with dissimilar attributes are positioned farther apart.

Scientists explore and compare ensemble members and data spatially within the ensemble view using two main interactions: observation-level interaction (OLI), subsetting of ensemble members, and ROI selection within ensemble members. OLI is an interaction technique built on the SI principles (Ender et al., 2012). It allows scientists to directly manipulate ensemble members to investigate and understand their commonalities and differences. The subsetting feature allows scientists to navigate subsets of ensemble members for a more in-depth analysis. This capability enables scientists to seamlessly switch between overview-first and detailed-first analysis modes. Moreover, scientists can utilize the ROI selection feature to identify specific spatial regions of interest within the data. This feature is valuable when dealing with a large grid where not all regions are equally significant to the analysis. This allows a focused exploration of regions with similar patterns or trends, enhancing the effectiveness of the analysis.

Parameter view (Figure 2b) offers scientists the opportunity to investigate parameter sensitivity using ensemble attributes (i.e., input parameters and outputs). Each attribute is represented on a horizontal slider. The slider's value represents the weight of the attribute in the model, thereby marking its importance. Using a Parametric Level Interaction (PLI), scientists are able to manipulate the slider in order to interact with model attributes. PLI allows scientists to investigate associations and relationships between input parameters and explore their impact on the simulation outputs.

Conversely, the statistical view (Figure 2c) helps find the optimal parameter settings. This view allows scientists to explore raw data through various statistical representations, such as parallel coordinates, scatterplots, and boxplots. By leveraging these different displays, scientists can determine variability in the data, validate conclusions, identify hidden correlations not found by other views, and gain insights into the distribution of different ensemble members.

4 SPATIAL ENSEMBLE MODELING

Simulation inputs and outputs serve as entry points to SpatialGLEE's visualization pipeline. Passing spatial ensemble raw data directly to the visualization pipeline would capture the spatiality in the data during the exploration and analysis. However, it poses computational challenges due to the complexity and size of the ensemble. Therefore, there is a need for modeling the spatial ensemble data while preserving the underlying spatial structure.

Given a multidimensional spatial ensemble $E = \{K_1, K_2, \dots, K_N\}$ with M members, where each $K_i \in E$ is of a 2D grid G . Every grid point within K_i is linked to input and output values obtained from spatial simulation assessments or measurements conducted across the entire grid. K_i represents a spatial stochastic process $\{Q(s) : s \in G\}$ with the spatial domain $G \subset \mathbb{R}^d$, $d \geq 1$ (d vector of coordinates)(Dahshan, 2021). We model $Q(s)$ by

$$Q(s) = X(s)\beta + w(s) + \epsilon, \quad (1)$$

so that $Q(s)$ has mean $X(s)\beta$ and error $w(s) + \epsilon$. The mean is the result of the product between the coefficients β and $X(s)$, which represents a vector of p co-variables at locations S . The additive error adds a spatial-dependent error term, $w(s)$, and an independent measurement error term, ϵ , characterized by a zero mean and variance τ^2 . The spatial dependence is imposed by modeling $w(s)$ as a stationary, mean-zero spatial process with a covariance function

$$C^S(s - s') = \sigma^2 g(\|\Sigma^{-1/2}(s - s')\|), \quad (2)$$

$g(\cdot)$ represents a spatial kernel covariance function that operates on the distance between two grid points, and σ^2 serves as a covariance inflation term. Spatial covariance kernel functions establish the characteristics and degree of spatial dependence within the spatial process. For instance, they can establish greater dependence between spatial outcomes when they are close compared to those at a greater distance.

Based on the discussions with our collaborators, we learned that their primary focus lies in understanding spatial autocorrelation and variability in the ensemble. According to Tobler's 1st Law of Geography (Miller, 2004), spatial autocorrelation expects nearby spatial grid points to be more similar than far apart ones. Therefore, our proposed approach fits GPR (Lázaro-Gredilla et al., 2010), also known as Kriging, to learn the characteristics of spatial processes. Kriging provides a method for interpolating between grid points. These interpolations maintain the spatial correlation and variability present in the data.

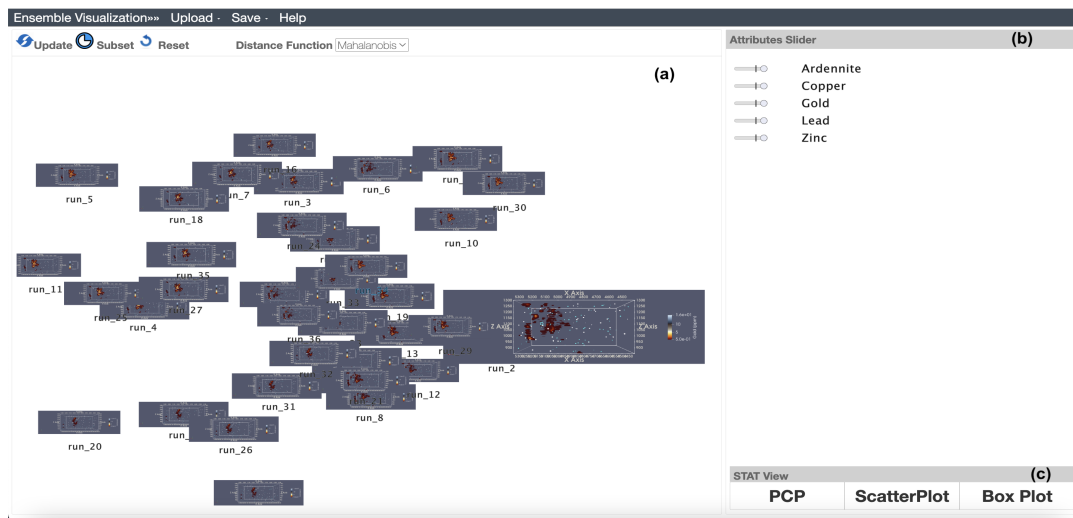


Figure 2: SpatialGLEE’s main interface:(a) *Ensemble view* shows the WMSD projection of the simulation ensemble in 2D space. Ensemble members are spatially arranged so that similar ensemble members are near each other and dissimilar ensemble members are far apart. Scientists can interactively manipulate image thumbnails representing ensemble members to explore and analyze them. (b) *Parameter view* presents the weights assigned to both input parameters and simulation outputs. Scientists can alter the slider values to examine the impact of the ensemble attributes. (c) *Statistical view* offers multiple statistical displays to explore and understand the distributions, patterns, trends, and outliers in ensemble raw data.

With simulation measurements $Y(s_i)$ of N grid locations s_i ($i \in \{1, \dots, N\}$). Estimating the process Y at a new location s_0 , denoted $Y^*(s_0)$ is a two-step process. The process begins by fitting a variogram to ascertain the spatial covariance structure and parameters based on observed data. Subsequently, it calculates the weights, denoted as λ_i , from the covariances between each observed location s_i and the new location s_0 . The value $Y^*(s_0)$ is then derived from a weighted average.

$$Y^*(s_0) = \sum_{i=1}^N \lambda_i Y(s_i). \quad (3)$$

In our approach, we model a second-order stationarity Gaussian Process (GP) with an isotropic Matern covariance kernel. (Nychka et al., 2002). The Matern covariance function is

$$C(d|\kappa, \nu, \sigma^2) = \sigma^2 \frac{2^{\nu-1}}{\Gamma(\nu)} (d/\kappa)^\nu K_\nu(d/\kappa) \quad (4)$$

where d is the Euclidean distance between s and s' ($\|s - s'\|$), $\Gamma(\cdot)$ is the gamma function, ν is a smoothing parameter that controls the mean square differentiability of the process ($\nu > 0$), κ is a range parameter, and $K_\nu(\cdot)$ is the modified Bessel function of second kind order. We selected the Matern covariance because it offers significant flexibility in modeling spatially correlated random processes. The smoothness parameter associated with Matern covariance allows control over the level of smoothness in the spatial process.

To estimate model parameters, Kriging maximizes the likelihood of the simulation measurements generating five estimates: nugget variance, scalar MLE for kappa, anisotropy parameters (λ_{m1} , λ_{m2}), and process variance. The computation of these estimates relies heavily on matrix operations. The distance matrices and covariance utilized in these operations capture the pairwise correlations among the locations of the grid. As the spatial grid size increases, the computational cost of kriging escalates, restricting its feasibility with large datasets. To tackle this challenge, various approaches have emerged, falling into two categories: sparse approximations (Kaufman et al., 2008) and local approximations (Wackernagel, 2013). Sparse approximations involve approximating the covariance matrix with sparse matrices. Conversely, local approximations partition input data into local and independent regions, with each spatial region encompassing a relatively small number of local points closest to the prediction point. Despite their ability to handle large datasets, these approximation approaches rely on some form of approximation, introducing the risk of information loss and errors, thereby limiting the advantages of using kriging.

Our aim is to implement a scalable high-performance parallel and distributed version of the MLE of the Kriging model. This implementation leveraged multi-core and multi-node architectures to boost computational performance. We leveraged the `Aniso_fit()` API from the `ConvoSPAT` package of `R` (Risser and Calder, 2015) for our implementation.

We modified the `Aniso_fit()` MLE implementation to evaluate the maximization function and the function returning the gradient for the same parameter value in parallel. We used the *Parallel* package from R, which supports multi-core and multi-node parallelization.

5 SPATIAL ENSEMBLE VISUAL EXPLORATION

SpatialGLEE's modeling and visualization aim to investigate similar behaviors and key parameters in the ensemble, aligning with design goals. Scientists can use SpatialGLEE's multi-linked views and interaction techniques to explore and analyze spatial ensembles. SpatialGLEE integrates scientists' intuition and expertise with machine learning and statistics, facilitating the analysis and exploration of multidimensional spatial ensembles. The visualization pipeline of SpatialGLEE is comprised of three main components: the input source, similarity models, and coordinated visual interfaces. The spatially estimated simulation outputs and input parameters serve as the input source, processed through similarity models for 2D projection and manipulation. Coordinated multi-views facilitate analysis, providing a comprehensive understanding of spatial relationships and patterns in the data.

5.1 Spatial Ensemble Attributes

The SpatialGLEE visualization pipeline employs the spatial estimates (i.e., anisotropy parameters (lam1 , lam2), nugget variance, process variance, and scalar MLE for kappa) as a foundation for spatial ensemble analysis and exploration. To ensure an accurate unbiased representation of the data during the exploration and analysis of the spatial ensemble, these estimates are z-score normalized. This normalization process helps standardize the data and ensures that each estimate is considered in the context of its distribution, preventing any skewed or misleading interpretations. In addition, an initial weight of $(1/d)$ is assigned to each attribute (i.e., inputs and outputs), where d is the total number of simulation outputs and input parameters. The weight of each attribute is evenly distributed among its estimates, resulting in a weight of $(1/5d)$ for each individual estimate.

5.2 Similarity Models

The similarity models in SpatialGLEE manage the visualization and interaction of simulation ensem-

ble data through two models: forward and backward. The forward model utilizes weighted multidimensional scaling (WMDS) to project multidimensional ensemble data into two-dimensional space. This weighted projection integrates a weighted distance function to capture commonalities and distinctions between ensemble members based on spatially estimated attributes. The determination of the distance function in SpatialGLEE depends on the data characteristics, the task, and the projection technique. The primary interface offers scientists a range of distance functions to choose from, including weighted Cosine, weighted Euclidean, and weighted Manhattan, with the default being the weighted Mahalanobis distance. The Mahalanobis distance measures the distance between any point in space and the distribution's center, taking into account correlations between attributes.

To explore ensemble members described by spatial estimates of d parameters, WMDS is applied using weighted distance $D_w(i, j)$, for ensemble members i and j ($i, j \in \{1, \dots, M\}$), with weight w_a representing the weight applied to each spatial estimate to denote its significance in the projection. The pairwise distance function result between ensemble members is subsequently inputted into WMDS. WMDS determines the position of each individual member in the low dimensions by minimizing the mean squared error between the pairwise distances in two dimensions and the corresponding distances in the high-dimensional space.

The backward model is activated when scientists interact with SpatialGLEE's different views using different interaction techniques (i.e., OLI, PLI, brushing, or subsetting). OLI enables scientists to create a customized spatialization of multidimensional ensemble members based on their intuition and domain expertise. For instance, the expertise of the scientists may contradict the spatialization of the ensemble members, or they may observe interesting patterns in the data. In response, they perform an OLI by dragging subsets of ensemble members into groups. These groupings signify scientists' hypothesized similarity between these ensemble members. The backward model is then triggered, calling upon a semi-supervised metric learning model. This model attempts to learn new weights that correspond to the identified similarity and subsequently adjusts the projection. Thus, OLI facilitates the exploration of relationships and associations between ensemble members by allowing scientists to shape the spatialization based on their insights and hypotheses.

PLI empowers scientists to explore parameter sensitivity by directly manipulating the attribute's weight

on the slider. This interaction results in an updated weight vector and a new projection of ensemble members based on the manipulated weight on the slider. Given that all weights are constrained to add up to one, increasing the weight of one attribute necessitates decreasing the weights of all other attributes, and vice versa. Thus, the updated projection amplifies the similarities and differences among data points, intensifying with an increase in weight and diminishing with a decrease in weight. This enables scientists to offer parametric feedback to the backward model regarding which attribute they consider to be significant and to observe how this attribute affects ensemble members' low-dimensional grouping. This interaction facilitates the exploration of the influence of individual attributes on the spatial representation, allowing scientists to gain insights into parameter sensitivity within the ensemble.

The average size of a spatial simulation grid may surpass millions of grid cells. In this case, scientists might be interested in examining specific regions, like those near wells or aquifers. The SpatialGLEE ROI selection feature enables the interactive selection of an ROI within the ensemble member within thumbnails in the ensemble view. Scientists can explore ROI in two ways: either exclusively focusing on an ROI or assigning a higher weight to it compared to other regions on the grid. Exploring a specific ROI across all ensemble members triggers the backward model, updating the visualization pipeline with a newly updated weight vector for spatial estimates of the chosen subregion of interest and thus adjusting the projection accordingly. On the other hand, increasing the weight of a particular ROI triggers the backward model, which updates the weight vector by dividing the weights between ensemble estimated attributes and subregion of interest estimated attributes based on a percent determined by the scientist.

5.3 Coordinated Multi-View Visualizations

SpatialGLEE's multi-coordinated views, Figure 2, contain an *ensemble view*, a *parameter view*, and a *statistical view*.

Ensemble View: visualizes multidimensional ensemble members in two-dimensional space using WMDS to provide an overview of spatial ensemble data. Each ensemble member is presented through a two-dimensional image of the simulation output. The ensemble view offers two main interactions: OLI and ensemble members and ROI selection. These two interactions provide scientists with an interactive exploration environment for analyzing simulation ensemble

members, thereby contributing to achieving the second and third design goals.

Parameter View: represents simulation input parameters and output parameters on horizontal sliders, with each attribute represented by a single slider whose value is the attribute's weight. The weight of each attribute is the sum of the weights of all its estimates. Scientists utilize PLI within parameter view to investigate parameter sensitivity, achieving the first design goal.

Statistical View: supports other views by allowing scientists to validate assumptions and refine hypotheses derived from other views. It offers three statistical displays—parallel coordinates, boxplot, and scatter plot—supporting univariate, bivariate, and multivariate analyses. It can also be used to gain insights into data distributions, identify the variability across different regions, discover new patterns undiscovered, or/and produce novel hypotheses that other views could confirm, thereby contributing to achieving the fourth design goal. Statistical View offers statistical displays for individual ensemble members, multiple ensemble members, and specific regions within ensemble members. For instance, scientists can explore correlations between ensemble members or analyze the distribution of single or multiple ensemble members across the entire grid or specific subsets.

6 EVALUATION

We evaluated our proposed approach using a 2D spatial ensemble from geologic CO₂ sequestration. During the evaluation, we examined the effectiveness of the proposed approach in aiding scientists to explore and analyze multidimensional spatial ensembles. Our emphasis was on gauging how well the approach facilitates the exploration of parameter sensitivity and optimization, as well as the examination of similarities and differences among ensemble members and spatial regions of interest. Specifically, we investigated the extent to which the statistically estimated parameters could preserve the spatial structure during exploration and analysis.

Three geoscience domain scientists (two graduate students and a faculty member) evaluated the proposed approach. The faculty member supplied the ensemble data used in the experiment. To initiate the evaluation, the scientists received instructions on how to use SpatialGLEE and its interaction techniques. Subsequently, they were tasked with utilizing SpatialGLEE's interaction techniques and visual interfaces to explore and analyze the ensemble. Throughout the evaluation, we recorded both the duration to complete

each task and the level of task completion. Additionally, we timed every interaction that the domain experts carried out.



Figure 3: The initial projection of the multidimensional CO2 flow ensemble using spatially estimated simulation attributes.

6.1 Case Study

The simulation ensemble study examines the multi-phase fluid dynamics of CO2 flow in a basalt fracture network (Gierzynski and Pollyea, 2017). The model domain has a 2-D fracture network based on high-resolution LiDAR scans of a basalt outcrop in the Columbia River Plateau. The 5 m x 5 m model domain is divided into 40,000 2.5 cm x 2.5 cm Cartesian grid cells. The study constructed a simulation ensemble that randomly assigns fracture permeability to each grid cell from a basalt core sample permeability distribution because centimeter-scale fracture permeability is unknown. Thus, the ensemble has 25 equally probable model domains with permeability spatial distribution as the random variable. Using e-type estimations, this simulation ensemble examined how permeability impacts buoyant CO2's flow characteristics during geologic CO2 sequestration in a basalt reservoir as it phases from supercritical fluid to subcritical gas.

Figure 3 shows the initial projection of a multidimensional CO2 flow ensemble in 2D space. The scientist didn't observe any interesting patterns or groups in the initial projection. The scientist was interested in understanding the parametric controls of the phase change to gas because the gas phase CO2 is significantly more buoyant and therefore has a greater chance of leaking out of the CO2 storage reservoir. To investigate this, the scientist started grouping ensemble members into two groups based on how much CO2 has moved to the upper portion of the model, which is where the CO2 undergoes a phase change to a gas phase, performing an OLI (Figure 4 a).

The reprojection of the ensemble revealed that there was a notable increase in the weight of gas-phase CO2, permeability, and fluid pressure, as il-

lustrated in Figure 4b. This led the scientist to conclude that fluid pressure is the dominant control for this grouping. The rationale behind this conclusion is that the phase change from supercritical CO2 to gas-phase CO2 occurs as the CO2 floats to shallower depths, where fluid pressure is lower. Additionally, the scientist gained insights from this reprojection that permeability affects CO2 flow; this necessitates additional experiments to determine how permeability distributions vary between ensemble members that exhibit phase change and those that do not.

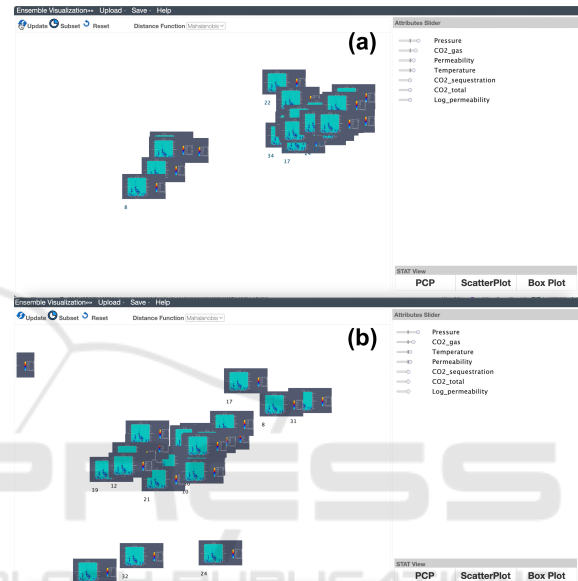


Figure 4: Investigating migration CO2 flow using OLI: a) semantically grouping ensemble members from the CO2 flow ensemble into two clusters based on how much CO2 has moved to the upper portion of the model; b) The resulted projection shows permeability and fluid pressure (P) have a dominant role in the phase change from supercritical CO2 to gas-phase CO2.

The scientist wanted to investigate the effect of fluid pressure alone on the ensemble, so s/he conducted PLI by increasing the fluid pressure's weight. However, the reprojection result was inconclusive. So, the scientist decided to explore if the temperature has any influence on the ensemble by performing PLI (Figure 5). The resulted projection was a linear projection in which ensemble members with high CO2 gas at shallow depths close to the top of the workspace. This discovery led the scientist to observe that there is a thermal control on the CO2 phase change, but it is much more subtle in the ensemble. This was a new discovery that required a more detailed analysis.

The scientist was interested in determining whether log permeability affects the ensemble. So,



Figure 5: Increasing the weight of the temperature attribute while performing PLI to explore its impact on the ensemble. The resulted projection led to the conclusion that there is a thermal control on the CO2 phase change.

s/he increased the weight of log permeability performing an PLI (Figure 6). The projection resulted in the separation of ensemble members into distinct piles. The pile on the right has low CO2 gas at shallow depths, whereas the pile on the left has high CO2 gas concentrations at shallow depths. This led the scientist to conclude that the pile on the right has low permeability in the conductive fracture shallow depth, which keeps fluid pressure high and prevents the CO2 from undergoing phase change to CO2 gas.

After interacting with SpatialGLEE through various interactions, the scientist aimed to identify potential patterns among parameters and investigate whether the distribution of raw data for different parameters varied across runs. To explore this, the scientist employed statistical view displays (Figure 7). The use of a boxplot showed that the distribution of log permeability remained consistent across all runs. Furthermore, a scatter plot uncovered a moderately positive correlation between pressure and CO2 sequestration, as well as a positive correlation between CO2 sequestration and CO2 total. By employing parallel coordinates, the scientist detected a correlation among log permeability, temperature, and CO2 gas across specific runs.

The scientist aimed to investigate regions of interest (ROI) within the reservoir. To facilitate this exploration, s/he resets the ensemble view to obtain a new projection. Subsequently, s/he selected a particular run and identified a ROI. SpatialGLEE supports two exploration options for ROI: either allocate a percentage of the weight vector to this ROI and explore it concurrently with the entire reservoir or assign the entire weight vector exclusively to this specific ROI. The scientist explored both options while using SpatialGLEE interaction techniques and determined that specific characteristics in these ROIs required further investigation through higher-fidelity simulations (Figure 8).



Figure 6: Performing a PLI to investigate the effect of increasing the log permeability weight on the ensemble. The resulted projection led to the conclusion that low permeability in the conductive fracture shallow depth keeps fluid pressure high and prevents CO2 from undergoing phase change to CO2 gas.

6.2 Domain Expert Evaluation

We discussed the SpatialGLEE tool and its interaction techniques with the domain experts who provided the ensemble data for the case study. We solicited their feedback on the usability and utility of SpatialGLEE. Scientists confirmed that by using SpatialGLEE, they were able to reach conclusions in significantly less time compared to the traditional analysis process. SpatialGLEE coordinated views allow them to simultaneously visualize, explore, and understand parameter and ensemble spaces in the absence of prior knowledge. The ensemble view allows analyzing both ensemble members and spatial regions within members. This helps them identify significant regions in the grid. The parameter view permits them to directly investigate input parameters' effects on simulation outputs. The statistical view supports scientists in locating interesting patterns in raw spatial data and exploring them using SpatialGLEE's views and interaction techniques. It also helps in finding the optimal parameter settings. SpatialGLEE has the potential to aid in the discovery of new insights that necessitate additional experiments for in-depth analysis.

6.3 Discussion

SpatialGLEE Showed the Potential to Improve Spatial Ensemble Exploration over Traditional Approaches. SpatialGLEE presents an approach for exploring multidimensional spatial ensembles when scientists do not have an in-depth understanding of the simulated model. Traditionally, scientists utilize visualizations of simulation outputs and summary statistics to investigate the variability of an individual parameter (whether input or output) throughout the entire ensemble. This traditional approach often ne-

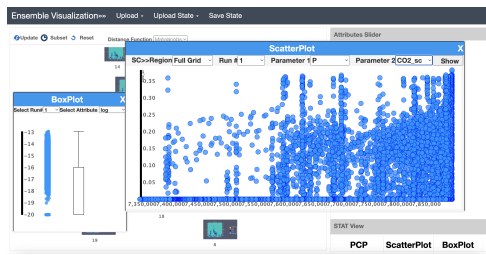


Figure 7: Exploring the distributions and correlations between parameters of raw spatial data utilizing the statistical view. This view shows statistical distributions and properties of data using univariate, bivariate, and multivariate statistical displays.



Figure 8: Spatial Region of Interest (ROI) Selection: a) The scientist was interested in exploring a certain region within the reservoir, so s/he selected the ROI and increased its importance over other regions in the grid by 100%. b) The ensemble members are re-projected based on the newly updated weight vector.

cessitates the implementation of several programs or scripts for data visualization, which takes a substantial amount of time and increases the risk of errors. In our comparison between the insights and conclusions generated by SpatialGLEE and those obtained through the scientist's regular analysis process, we found that SpatialGLEE led to the same conclusions as the manual analysis process but in significantly less time. Furthermore, SpatialGLEE facilitated the discovery of new phenomena and insights that would be challenging to uncover using traditional methods.

The qualitative analysis of the study reveals that using OLI, the scientist was able to figure out commonalities and differences across ensemble members and within ROI in ensemble members based on specific patterns or hypotheses. The resulted projection from OLI provides scientists with the ability to identify which parameter(s) guide the grouping of ensemble members. However, OLI is a technique for exploratory interaction; therefore, it would not always generate significant outcomes. Obtaining

significant insights or even discoveries would be possible if the grouped ensemble members shared high-dimensional features that could be captured by the metric learning model. Utilizing the PLI, scientists managed to identify the sensitivity of input parameters to the simulation output. This facilitated the identification of crucial parameters and those that could be set as constants in the simulation. On the other hand, through the use of statistical view displays, scientists were able to explore and analyze raw data from spatial ensembles. This helped them understand the distribution and variability of the data, leading to the identification of optimal parameter settings for the input parameters.

SpatialGLEE Performance. Our quantitative measurements demonstrate that the interaction techniques of SpatialGLEE allow scientists to accomplish all preliminary exploration tasks. However, when interacting with SpatialGLEE, additional exploratory inquiries emerged. The scientists addressed some of these inquiries, but others demanded simulations of greater fidelity. We computed the total number of interactions needed to answer the preliminary exploration tasks. This count varied among scientists, depending on the nature of the interactions they performed within the ensemble. For example, completing the preliminary exploration tasks needed on average took 1–5 interactions. Advanced tasks that emerged during the analysis exhibit variability in the number of interactions, ranging from an average of 3 to 8. Additionally, SpatialGLEE responded to the interactions of scientists within a reasonable time frame. Responding to OLI, PLI, and ensemble member and ROI selection required less than 5s, 3s, and 1s respectively.

Spatial Ensemble Modeling. Modeling spatial ensembles using GPR preserves the spatiality of the data during exploration and analysis of both ensemble members and ROI within the ensemble. Using the modeled data with SpatialGLEE interaction techniques revealed that grouping ensemble members captured the intrinsic structures and spatial characteristics of the data. This provided scientists with additional insights that might be challenging to obtain using traditional visualization methods reliant on summary statistics (i.e., standard deviation and mean). Moreover, the parallel implementation of MLE demonstrates superior computational performance compared to conventional MLE. Our approach is scalable, functioning effectively across both the same node and multiple nodes (Figure 9). We conducted our scaling evaluation on an Intel SkyLake Xeon Gold cluster with 24 cores and 384 GB of mem-

ory per node. We observed that while our implementation scales on multiple cores of the same node, using multiple nodes leverages the aggregate cluster memory, resulting in further performance improvement. On 8 nodes and 128 cores, our parallel implementation achieved approximately $8\times$ speedup compared to 2 nodes and 16 cores. The total speedup compared to the sequential `Aniso.fit()` implementation is $21\times$.

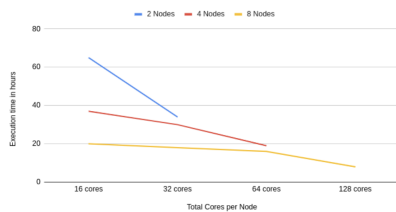


Figure 9: Execution time of parallel MLE on various cores (i.e., 16, 32, 64, 128) across distributed nodes (i.e., 2, 4, 8)

Distance Function Learning Model. SpatialGLEE is designed to assist scientists in exploring spatial simulation ensembles. The projection of ensemble members from high-dimensional space to 2D space and their grouping in the ensemble view significantly influence the exploration process. In the forward similarity SI model, the distance function can be described as an interactive distance function learning model. The selection of the distance function plays a pivotal role in the forward similarity SI model and, consequently, impacts the projection of ensemble members. The interactive nature of the distance function suggests that it may learn from the specific characteristics of the data, contributing to a more dynamic approach to determining similarities between the ensemble members. While the choice of distance function is based on the characteristics of the data and the task at hand, we observed different performances of different distance functions (i.e., Euclidean, Manhattan, and Mahalanobis) when using spatial ensembles. The Euclidean distance only captures the general spatial arrangement in the ensemble, providing limited insights to scientists. Conversely, Manhattan distance outperforms Euclidean distance due to its sensitivity to multivariate outliers. In contrast, when compared to other distance functions, Mahalanobis distance demonstrates greater accuracy in grouping ensemble members. This can be attributed to its consideration of the multivariate covariance structure during distance calculation. However, Mahalanobis distance comes with a notable drawback: its computational complexity is high for large datasets.

Limitations and Future Work. Increasing the number of ensemble members to the hundreds may result

in visual cluttering within the ensemble view. One potential solution to the problem is using larger displays that are capable of accommodating a greater number of ensemble members. Based on scientists' feedback, it was observed that scientists typically opt for an ensemble size that is smaller than one hundred. Our current approach and its parallel implementation are tailored for 2D spatial grids and do not currently provide support for 3D grids. In our future work, we plan to expand SpatialGLEE to add support for 3D grids and incorporate the capability for handling time-varying simulation ensembles.

7 CONCLUSION

In this paper, we proposed SpatialGLEE, a visual exploration approach for multi-dimensional spatial ensembles. The proposed approach modeled the spatiality of data in ensemble members using Gaussian Process Regression (GPR) and explored its feasibility for visual exploration with Semantic Interaction. SpatialGLEE interactive visual interfaces and interactions enabled scientists to explore commonalities and distinctions across ensemble members, subsets of the ensemble, and ROI within the ensemble. Additionally, they were able to determine parameter sensitivity and optimization, as well as analyze the static properties of the raw spatial data of ensemble members and their parameters. The effectiveness of our proposed approach was evaluated through experiments involving domain experts. We found that by employing the SpatialGLEE approach, scientists could effectively explore spatial simulation parameter and ensemble spaces simultaneously, potentially leading to the generation of new findings and discoveries.

REFERENCES

- Ahmed, K., Sachindra, D. A., Shahid, S., Demirel, M. C., and Chung, E.-S. (2019). Selection of multi-model ensemble of general circulation models for the simulation of precipitation and maximum and minimum temperature based on spatial assessment metrics. *Hydrology and Earth System Sciences*, 23(11):4803–4824.
- Athawale, T., Maljovec, D., Yan, L., Johnson, C., Pascucci, V., and Wang, B. (2020). Uncertainty visualization of 2d morse complex ensembles using statistical summary maps. *IEEE Transactions on Visualization and Computer Graphics*.
- Cappello, F., Constantinescu, E., Hovland, P., Peterka, T., Phillips, C., Snir, M., and Wild, S. (2015). Improving the trust in results of numerical simulations and scientific data analytics. Technical report, Argonne National Lab.(ANL), Argonne, IL (United States).

- Chen, X., Shen, L., Sha, Z., Liu, R., Chen, S., Ji, G., and Tan, C. (2019). A survey of multi-space techniques in spatio-temporal simulation data visualization. *Visual Informatics*, 3(3):129–139.
- Dahshan, M., Polys, N. F., Jayne, R., and Pollyea, R. M. (2020). Making sense of scientific simulation ensembles with semantic interaction. In *Computer Graphics Forum*, volume 39, pages 325–343. Wiley Online Library.
- Dahshan, M. M. S. I. (2021). Visual analytics for high dimensional simulation ensembles.
- Demir, I., Dick, C., and Westermann, R. (2014). Multi-charts for comparative 3d ensemble visualization. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2694–2703.
- Ender, A., Fiaux, P., and North, C. (2012). Semantic interaction for sensemaking: inferring analytical reasoning for model steering. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2879–2888.
- Evers, M. and Linsen, L. (2022). Multi-dimensional parameter-space partitioning of spatio-temporal simulation ensembles. *Computers & Graphics*, 104:140–151.
- Favelier, G., Faraj, N., Summa, B., and Tierny, J. (2018). Persistence atlas for critical point variability in ensembles. *IEEE transactions on visualization and computer graphics*, 25(1):1152–1162.
- Fofonov, A. and Linsen, L. (2018). Multivisa: Visual analysis of multi-run physical simulation data using interactive aggregated plots. In *VISIGRAPP (3: IVAPP)*, pages 62–73.
- Fofonov, A. and Linsen, L. (2019). Projected field similarity for comparative visualization of multi-run multi-field time-varying spatial data. In *Computer Graphics Forum*, volume 38, pages 286–299. Wiley Online Library.
- Gierzynski, A. O. and Pollyea, R. M. (2017). Three-phase co2 flow in a basalt fracture network. *Water Resources Research*, 53(11):8980–8998.
- Huang, Q., Chen, Q., Liu, G., and Cui, Z. (2023). Visualization facilitates uncertainty evaluation of multiple-point geostatistical stochastic simulation. *Visual Intelligence*, 1(1):12.
- Huesmann, K. and Linsen, L. (2022). Similaritynet: A deep neural network for similarity analysis within spatio-temporal ensembles. In *Computer Graphics Forum*, volume 41, pages 379–389. Wiley Online Library.
- Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555.
- Kumpf, A., Stumpfegger, J., Hartl, P. F., and Westermann, R. (2021). Visual analysis of multi-parameter distributions across ensembles of 3d fields. *IEEE Transactions on Visualization and Computer Graphics*.
- Lázaro-Gredilla, M., Quinonero-Candela, J., Rasmussen, C. E., and Figueiras-Vidal, A. R. (2010). Sparse spectrum gaussian process regression. *The Journal of Machine Learning Research*, 11:1865–1881.
- Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., and Di, Z. (2017). A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdisciplinary Reviews: Water*, 4(6):e1246.
- Liu, L., Padilla, L., Creem-Regehr, S. H., and House, D. H. (2018). Visualizing uncertain tropical cyclone predictions using representative samples from ensembles of forecast tracks. *IEEE transactions on visualization and computer graphics*, 25(1):882–891.
- Miller, H. J. (2004). Tobler’s first law and spatial analysis. *Annals of the Association of American Geographers*, 94(2):284–289.
- Mirzargar, M., Whitaker, R. T., and Kirby, R. M. (2014). Curve boxplot: Generalization of boxplot for ensembles of curves. *IEEE transactions on visualization and computer graphics*, 20(12):2654–2663.
- Nychka, D., Wikle, C., and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2(4):315–331.
- Orban, D., Keefe, D. F., Biswas, A., Ahrens, J., and Rogers, D. (2018). Drag and track: A direct manipulation interface for contextualizing data instances within a continuous parameter space. *IEEE transactions on visualization and computer graphics*, 25(1):256–266.
- Potter, K., Wilson, A., Bremer, P.-T., Williams, D., Doutriaux, C., Pascucci, V., and Johnson, C. R. (2009). Ensemble-vis: A framework for the statistical visualization of ensemble data. In *2009 IEEE International Conference on Data Mining Workshops*, pages 233–240. IEEE.
- Risser, M. D. and Calder, C. A. (2015). Local likelihood estimation for covariance functions with spatially-varying parameters: the convospat package for r. *arXiv preprint arXiv:1507.08613*.
- Shi, N., Xu, J., Li, H., Guo, H., Woodring, J., and Shen, H.-W. (2022). Vdl-surrogate: A view-dependent latent-based model for parameter space exploration of ensemble simulations. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):820–830.
- Shu, Q., Guo, H., Liang, J., Che, L., Liu, J., and Yuan, X. (2016). Ensemblegraph: Interactive visual analysis of spatiotemporal behaviors in ensemble simulation data. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pages 56–63. IEEE.
- Vietinghoff, D., Bottinger, M., Scheuermann, G., and Heine, C. (2022). Visualizing confidence intervals for critical point probabilities in 2d scalar field ensembles. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pages 145–149. IEEE.
- Wackernagel, H. (2013). *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media.
- Wang, J., Hazarika, S., Li, C., and Shen, H.-W. (2018). Visualization and visual analysis of ensemble data: A survey. *IEEE transactions on visualization and computer graphics*, 25(9):2853–2872.
- Wang, J., Liu, X., Shen, H.-W., and Lin, G. (2016). Multi-resolution climate ensemble parameter analysis with nested parallel coordinates plots. *IEEE transactions on visualization and computer graphics*, 23(1):81–90.
- Winsberg, E. (2013). *Computer simulations in science*.
- Zhang, M., Chen, L., Li, Q., Yuan, X., and Yong, J. (2020). Uncertainty-oriented ensemble data visualization and exploration using variable spatial spreading. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1808–1818.