INTERACTIVE EXPLORATORY VISUAL ANALYTICS APPROACH FOR DISTRIBUTED SPATIOTEMPORAL DATA

by

Abdullah-Al-Raihan Nayeem

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing & Info Systems

Charlotte

2023

Approved by:

_____

Dr. Isaac Cho

_____

Dr. Zachary Wartell

_____

Dr. Wenwen Dou

_____

Dr. Huikyo Lee

_____

Dr. Mohamed Shehab

ABSTRACT

ABDULLAH-AL-RAIHAN NAYEEM. Interactive exploratory visual analytics approach for distributed spatiotemporal data. (Under the direction of DR. ISAAC CHO)

Spatiotemporal visual analysis research is rapidly emerging and evolving, as scientists engaging in cross-domain research efforts, introducing novel systems to perform data-driven spatiotemporal analysis in several domains such as astronomy, climatology, environmental science, urban planning, etc. However, due to the increasing volume and complex nature of the spatiotemporal data, scientists encounter challenges to hypothesize, investigate, and compare the spatiotemporal variables. Often the spatiotemporal data are stored in disparate and fragmented forms in distributed data sites. Therefore, important insights may not reside in a single spatiotemporal dataset but rather distributed in multiple remote data sites. Such a scenario introduces a massive overhead in terms of acquiring the data, preparing the computational environment, and conducting exploratory spatiotemporal analysis. Moreover, our literature review and surveying the domain researchers suggest that exploratory visual analysis of spatiotemporal data still significantly relies on static visualizations to present specific data stories. An interactive visual analytics system obtain the potential to benefit in this context, providing a dynamic platform for scientists in performing distributed spatiotemporal data exploration.

In this dissertation, we address these challenges to outline the design requirements for a cloud-based visual analytics approach that serves distributed spatiotemporal data exploration. The interactive exploratory spatiotemporal visual analytics approach consists of three major components. First, we present a distributed data mining architecture in a visual analytics framework that supports unified spatiotemporal data access, transformation, and analysis. Next, we present an interactive contour-based geospatial visualization that supports exploratory and comparative

geo-spatiotemporal visual analysis. Finally, we present a pipeline for a visual analytics interface that sources distributed spatiotemporal data using the data mining architecture. To support the scientists in exploratory analysis, the pipeline provides interactive contour visualization in coordinated multi-views. To demonstrate the scalability and scientific value of the analytical workflows, we conducted qualitative and quantitative user studies. Results from the user intervention study and domain experts' feedback suggest that the proposed interactive visual analytics approach significantly improves users' performance in performing exploratory analysis over the distributed spatiotemporal data.

DEDICATION

To Aronnok and Neeladri,

My sweet and adorable nephew and niece.

ACKNOWLEDGEMENTS

First and foremost, I thank The Almighty Allah for giving me the strength, endurance, and perseverance that gets me to this day. I always find myself privileged, having a remarkable group of people in my life. These people always backed me up at their best to inspire me to achieve beyond my grasp. I want to acknowledge them here, as difficult as it is to put my wholehearted gratitude into words.

I want to pay heartfelt thanks to my supervisor Dr. Isaac Cho. He has been an excellent supervisor and mentor. His constructive feedback always helped to improve my research efforts. I am grateful for his active involvement and continuous guidance toward this dissertation. I would also like to recognize Dr. William J. Tolone for being an excellent PI and for his admirable efforts in providing the resources and tools for us to succeed. Thanks to Dr. Huikyo Lee for the research and collaboration opportunities with NASA JPL and his valuable research directions.

I am grateful to every member of our research group, especially, Dongyun Han, Donghoon Kim, Todd Dobbs, and Mohammed Elshambakey for their generous contributions to my research. I also want to thank all the members and collaborators of the Ribarsky Center for Visual Analytics. I feel proud to be a part of this research lab. Moreover, my gratitude to Dr. Zachary Wartell, Dr. Isaac Cho, Dr. Wenwen Dou, Dr. Huikyo Lee, and Dr. Mohammed Shehab for their valuable time serving on my dissertation committee.

I acknowledge the financial support by UNC Charlotte and the College of Computing and Informatics through the Graduate Assistant Support Plan (GASP), Graduate School Summer Fellowship, and William Ribarsky Memorial Scholarship.

My journey would not have been possible without the support of all my family members and friends. While my name is on the dissertation, every one of them has a fair share of credit for my achievements. I thank my parents, Mohammed Nuruzzaman and Rokeyza Zaman, for their life-long sacrifices. My brother Nahian,

the best anyone can ask, has always inspired me. I am eternally grateful for all his selfless support through thick and thin. I have never been close to being an academic excellence. Despite that fact, they never stopped giving me positive encouragement, even when they had every reason to act otherwise. I am forever grateful to them for that.

I could not have asked for better friends than Bivor, Moon, Chondon, and Putul to accompany me on this journey. These people have never failed to turn my most stressful days into colorful memories. My heartfelt thanks to Moon for being such a kind, caring friend and always being there in my ups and downs. My gratitude to Aquib and Tasmia for always being amazing friends to me. All the times I "wasted" hanging out, sharing laughs, and having late-night chats with all these people have only encouraged me to work harder the next day.

Last but not least, I am thankful to my wonderful wife, Fatema-E-Jannat Putul, for her unwavering love, patience, and encouragement in this journey. I would have neither pursued nor been able to complete this degree if it were not for her. I cannot emphasize her contribution enough to my research with invaluable suggestions and feedback. I have come this far for all the good she sees in me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANOVA An acronym for Analysis of Variance.

API An acronym for Application Programming Interface.

AWS An acronym for Amazon Web Services.

CI An acronym for Confidence Interval.

CMIP An acronym for Coupled Model Intercomparison Project.

CMV An acronym for Coordinated Multiple Views.

DAS An acronym for Distributed Analysis System.

DCP An acronym for Downscaled Climate Projection.

EBM An acronym for Error Bound Mean.

MTurk An acronym for Amazon Mechanical Turk.

NCA An acronym for National Climate Assessment.

NetCDF An acronym for Network Common Data Form.

NEX An acronym for NASA Earth eXchange.

SD An acronym for Statistical Downscaling.

UI An acronym for User Interface.

VAF An acronym for Visual Analytics Framework.

VIFI An acronym for Virtual Information Fabric Infrastructure.

CHAPTER 1: INTRODUCTION

In recent years, visualization researchers have made significant advances in spatiotemporal visual analysis. Interactive visualizations, today, enable scientists to hypothesize, investigate, and compare spatiotemporal variables in innovative ways leading to deeper, more insightful understandings. However, the breadth of earth science and its diversity with numerous branches and sub-disciplines [5, 6] demand visual analytic capabilities that can meet diverse sets of requirements [7, 8]. To create cognitively efficient visual analytic solutions for spatiotemporal data - data that are frequently high dimensional and complex - researchers often propose use-case specific approaches for visualizations and user interactions [9]. One emerging subarea of spatiotemporal analysis that embodies this characteristic is the visual exploration of spatiotemporal data from distributed data sources.

To support decision-making in a spatiotemporal data-driven society, researchers seek to exploit the power of big data and the benefits of derived insights, scientific discoveries, and enhanced understanding. The advance and convergence of methods and technologies - including advances in machine learning and deep learning methods; increased storage capacities and reduced storage costs; higher network speeds and larger network bandwidth; more economical and powerful high-performance computing; and a growing prevalence of sensor networks and smart technologies - are essential enablers to enhanced sense-making over big data. However, it is often the case for spatiotemporal data that important insights and discoveries do not reside within a single dataset, but instead are embedded within and across multiple and distributed datasets. Therefore, utilizing the maximal potential for data-driven insights necessitates analyses and sense-making, that occur across these distributed, disparate datasets - analyses and sensemaking that, thereby, enable accurate and reliable revelation of latent, complex correlations, patterns, relationships, and such other knowledge that may not be revealed from a single dataset alone.

Spatiotemporal data illustrates spatial change over time, studied in several research domains such as astronomy [10, 11], environmental science [12, 13, 14], urban planning [15], medical science [16], etc. Climate science is also discovered as a concentration that produces a wealth of complex, high dimensional data stored in a distributed environment which are critical for understanding climate change and its social impacts, as well as for effective decision-making for adaptations. Climate scientists are continuously assessing the spatiotemporal climate variables such as precipitation, temperature, and air quality to develop global climate models [17, 18, 19]. Exploratory visual analytics systems have the potential to significantly reduce [20] the burden of traditional spatiotemporal analysis workflows for climate scientists. In addition, technology and infrastructure advancements are providing greater public access to climate projection data [21]. In fact, researchers today can access climate data in distributed analytic environments [22] and render exploratory visualizations for analyses. Scientists are also working to optimize the computational efficiency of these analyses to enable real-time exploration of spatiotemporal data [23, 24]. These advances unfold opportunities for us, the visualization researchers to innovate over the full landscape of challenges and requirements to support interactive, in-situ visual analysis [25, 26] for the exploration of spatiotemporal climate data.

1.1 Thesis Statement & Contributions

In spatiotemporal analysis, data movement and variable extractions come at the expense of huge memory capacity, computational resources, and processing time. Advancements in technology and infrastructure over the last decade help to acquire more granular levels of spatiotemporal data. Earth science and geographical scientists are tirelessly working to develop novel analyses that help understand the changes in climate, their potential impact on the earth, and mitigation strategies. We have conducted a comprehensive literature review, domain expert survey, and collaboration with climate scientists to identify a few research gaps in the distributed spatiotemporal

analysis which we can contribute with an exploratory visual analytics approach. We outline the thesis statements based on finding answers to the following questions.

1. **Can a cloud-based architecture support exploratory spatiotemporal data analysis?**

   We hypothesize that a cloud-based architecture can facilitate data preparation without the overhead of massive data downloads. Additionally, it can enable standardized accessibility of massive data, mediating redundant user interaction with the distributed servers.

2. **Can a web-based visual analytics system support analysis of distributed spatiotemporal data?**

   We hypothesize that a web-based visual analytics approach can enable scientists and analysts to visually explore and compare spatiotemporal data. The interactive visualizations outlined in the visual analytics interface can facilitate the exploration and logical reasoning of the primary and resulting data. Moreover, domain researchers working in a common platform can enhance collaboration and foster knowledge sharing among the research community.

3. **Can interactive and coordinated multi-views support the users in visually performing their spatiotemporal analysis tasks?**

   We hypothesize that comprehensive user interactions and coordinated multiple views can reduce the completion time as well as the accuracy of estimation while performing the exploratory analytical tasks. These interactive visualization techniques can cater to the scientists and researchers in identifying not only the trends and unknown patterns in the spatiotemporal data but also the structural information and characteristics of the disparate and fragmented datasets.

In this dissertation, we address the research gaps in spatiotemporal data analysis and interactive visual exploration. We present our supporting research efforts,

comprehensive literature reviews, and obtained domain knowledge in developing exploratory visual analytics systems for distributed spatiotemporal data. In turn, we demonstrate a visual analytics approach that supports the scientists in performing their exploratory analysis tasks with distributed spatiotemporal data. This dissertation presents three major components - data mining architecture facilitating a framework for visual analytics systems, a contour-based interactive geospatial visualization, and a visual analytics interface.

1.2 Dissertation Outline

This dissertation is organized as follows. Chapter 2 summarizes related works in spatiotemporal visualization approaches and distributed visual analytics systems. We studied the data structures, storage formats, and exploratory analysis approaches that uncover insights from disparate data sources. Moreover, we discuss the recent upward trends of interactive visualizations in analyzing spatiotemporal data.

In chapter 3, we present a visual analytics framework (VAF), that facilitates distributed spatiotemporal data analysis in a visual analytics environment. We present usage scenarios from Sustainable Human Building Ecosystem (SHBE) and climate science research domains to establish the scientific value of the framework. Chapter 3 also includes a distributed data mining architecture that addresses the research problem of assisting scientists and researchers with their data acquisition, extraction, and scientific analysis tasks.

In chapter 4, we present an interactive geospatial visualization that facilitates exploratory analysis of the geospatial data. We demonstrated a set of interactive features, developed to cater to traditional geospatial visualization. In addition, we report a quantitative evaluation of the interactive geospatial view compared to the traditional view in performing seven exploratory analytical tasks including data association, clustering, ranking, etc.

In chapter 5, we illustrate the research effort in developing a visual analytics system

for distributed spatiotemporal data where we provide three use cases for the downscaled climate projections over the contiguous US. We leverage the distributed VAF in Chapter 3 to illustrate a pipeline for the visual analytics system for large-scale spatiotemporal datasets. In addition, we present coordinated multiple views (CMV) and interactive visual workflow for exploratory spatiotemporal analysis.

This dissertation contributes to distributed spatiotemporal data analysis, an objective that corresponds to multiple research domains such as atmospheric science, meteorology, political science, urban planning, and the human-building ecosystem. It provides a platform to learn underlying knowledge in disparate and fragmented data, extract insights, and collaborate with domain scientists and researchers. This interdisciplinary work includes comprehensive background research, to design scalable interactive pipelines that involve cross-domain experts, and the implementation of a visual analytics approach that facilitates the exploratory analysis over distributed spatiotemporal data.

CHAPTER 2: BACKGROUND & LITERATURE REVIEW

In this chapter, we provide a survey of literature on geo-spatiotemporal data mining and visual analysis. The structure of the literature review is as follows: section 2.1 provides an overview of distributed spatiotemporal data and its various transformations. We also reflect on different aspects of storing and encoding large-scale spatiotemporal data sets. In section 2.2, we discuss analytics and applications of spatiotemporal data sources. In turn, we introduce the analysis techniques and different applications of distributed spatiotemporal variables. In section 2.3, we survey the visual analytics systems for the geo-spatiotemporal data. We provide a comprehensive review of the visualization techniques and corresponding user interactions employed for spatiotemporal visual analyses. Finally, in section 2.4, we discuss the limitations of the existing systems in the exploration and comparative analysis of the distributed geo-spatiotemporal data for identifying trends and patterns.

2.1 Spatiotemporal Data

In this section, we present the definition of spatiotemporal data and its different transformations. To contextualize the definition, we often leverage climate science as a scenario where we scope the exploratory spatiotemporal analyses that are employed to study climate change, its impact and possible adaptations.

2.1.1 Data Definition

Events in spatiotemporal data consist of space and time. In simple terms, spatiotemporal data denotes events related to location that occurs over time. Data items related to a location or space are the geospatial or spatial data, traditionally represented by raster data, vector data or network data [9]. Raster data encodes a collection of pixels that each pixel represents a specific geographic location. This form of continuous spatial data illustrates information related to climate science (e.g., temperature, precipitation, and elevation). In contrast, vector data is represented by

Figure 2.1: A conceptual illustration of different data types classified in [1]. He et al. [2] illustrated the concept of the data type classification: (a) Multidimensional: 0D, 1D, 2D, and 3D; (b) Multivariate: scalar, 2-tuple, and n-tuple. (c) Network and (d) Tree data types are illustrated from the view diagram presented by Tominski et al. [3].

a specific point, line or polygon to denote a geographical area such as region, state, and building. The spatial network data generally represents road transport and its connected area where each edge may point to a spatial raster. The spatial data, on the other hand, is often vary on the time and tagged with a timestamp, called spatiotemporal data. The application of spatiotemporal data is found in many research domains such as climatology [21, 13], astronomy [10, 11], medical science [16], urban planning [15], etc.

2.1.2 Data Types

Research domain discussed above produce diverse spatiotemporal data based on the dimensions, events and structure. We categorize different data types and sources employed in the visual analytics interfaces according to the data type taxonomy discussed by Shneiderman [1]. This taxonomy was later utilized by Afzal et al. [27] for the ocean and atmospheric datasets. The data type taxonomy proposed by Shneiderman includes 1-2-3 dimensional, temporal, multivariate, tree, and network data [1]. He et al. [2] presented a conceptual illustration of multidimensional and multivariate data among other types as shown in Fig. 2.1a and b. In addition, we leveraged the data representation diagram presented by Tominski et al. [3] to illustrate the tree and network data types as shown in Fig. 2.1c and d.

1-Dimensional: This data type represents linear data such as textual documents,

a list of names, and source codes. The data are organized in a sequential layout, generally consisting of unstructured information. As a result, it is particularly challenging to design visualization for this data type. We did not find any instance of this type in context to spatiotemporal variables.

2-Dimensional (2D): This data type represents the planar data (i.e., x and y) that includes a geographic map, floorplans, or layouts. Datasets in this type generally have additional attributes such as name, value, etc. 2D climate data often represents the spatial distribution of a variable [28, 14]. Visual analytics interfaces often visualize these datasets with a multi layer approach on a 2D map. The user can toggle the layers of the map based on the variable of interest [29, 30, 31].

3-Dimensional (3D): This data type includes 3D models of real-world objects (e.g., buildings, Computer Aided Design (CAD) models) and scientific datasets (e.g., computed tomography scan). The data often obtain attributes with volume and complex interrelation with other data items. Volumetric rendering is a popular technique to visualize 3D scientific data [32, 33, 34]. Most modern climate models produce 3D data to study cross-correlation map among climate variables [35]. Climate change impact and adaptions related analysis also produce this type of data [36].

Temporal: This data type denotes the datasets that obtain time-varying variables. Temporal data denotes the changes in climate variables over time such as temperature, precipitation, humidity, etc. Data sources utilized in most of the visual analytics interfaces discussed in this paper relate to this data type [24] as temporal dimension obtains an essential portion of spatiotemporal data. Temporal data generally visualized using a standard 2D technique such as area plot [37], scatter plot [38], animation [39], timesliders [40, 41], etc. based on the data transformation.

Multi-dimensional: This data type includes datasets with multiple dimensions and attributes. Multidimensional data denotes multiple dimensions in the dataset whereas multivariate data denotes the multiple attributes associated with each di-

mension. Scientists and researchers study a wealth of multivariate data [3] to analyze variable association [36], distribution [42], path ensemble [14], etc. Multidimensional data is usually visualized using different 2D techniques such as integrated multiple visualization setup [43], small multiples [42, 44], color/glyph stylization [45, 46], or summarized form with fewer dimension [47]. Parallel coordinates are widely used for visualizing multidimensional and multivariate dataset [36, 14].

Tree: This data type represents the hierarchically structured data where each item in the collection has a link to the parent item except for the root. The data items are often linked based on multiple attributes. These datasets are visualized using the different representation of tree/hierarchical visualizations (e.g., edge-bundling, treemap). Kappe et al. [48] detailed a hierarchical clustering based on the climate states and time-dependent ensembles. They used dendogram and hierarchical clustering to visualize the data. In another work, we summarized the spatiotemporal precipitation and temperature data to produce hierarchical seasonal-regional mean [44]. However, we understand that tree structured data are not very common for spatiotemporal analyses as we found only a few examples of this data type.

Network: This data type includes the links among the data attributes that cannot be represented using a tree structure. Network data items often obtain an arbitrary number of relationships with other items that are commonly visualized using a node-link diagram and matrix representations. Kalo et al. [24] studied spatiotemporal interpolation of air pollution based on the data collected at measurement sites across the United States. They used a node-link diagram to visualize the triangulation of air pollution data. Analysis of path ensemble [49], spatiotemporal trajectory [50], variable similarity [42], data flow [51] etc. derive network data for climate variables.

2.1.3 Data Storage

Storing the spatiotemporal data has been a challenge for the analyst and research community due to its volume and complex nature [52]. Spatiotemporal data are often

characterized as complex because of high dimensionality, mutlivariability, and multi resolution factors [36]. Advancement in infrastructure and equipment has made it possible to acquire a more granular level of spatiotemporal data, the demands vast and faster storage units. For faster transaction, spatiotemporal data are frequently stored in a structured relational database (e.g., PostgreSQL [53], PostGIS [54]). However, the large volume of data from multiple sources is often convenient to store in a non structured format, known as NoSQL (e.g., Redis [55], MongoDB [56]). Storage is still a concern for spatiotemporal data for efficient data encoding and retrieval.

In climate science, the increasing resolution of the geospatial data and the frequent temporal records makes Network Common Data Form (NetCDF) [57], a well accepted data storing format in the research community for multidimensional spatiotemporal variables such as temperature, precipitation, and air quality. [58, 52]. NetCDF provides an encoding convention for the researchers to store the multidimensional data in a relatively smaller space [59]. Moreover, NetCDF allows appending streaming data without copying or defining the structure from the scratch. Individual data files are simultaneously accessible by one writer and multiple readers [57]. Climate scientists have established metadata conventions for storing climate variables in NetCDF file [60]. The conventions include the acceptable data types, naming convention, dimensions, unit of measurement, etc. The metadata container should also describe the title, data source, institute, and necessary references for the data [60]. Hence, NetCDF becomes one of the widely used formats for storing earth science data these days [61].

Besides, based on the general use case and characteristics of the data, spatiotemporal datasets are also stored as fragmented, partitioned clusters in distributed servers [51, 44]. As a result, the exploratory analysis performed by researchers and scientists does not always reside in a single data server. Important insight and knowledge base often reside in distributed and disparate datasets. Spatiotemporal data frag-

mented across multiple remote data hosts are identified as distributed spatiotemporal datasets. Big data computing infrastructure such as Hadoop supports large-scale data encoding in Hadoop Distributed File System (HDFS) [62] to enable distributed analysis [39].

2.2 Exploratory Analysis Tasks

Diversified exploratory tasks are employed to scrape the underlying insights from the distributed spatiotemporal data in context to the research domains. In this section, we contextualize the exploratory spatiotemporal analytical tasks in terms of climate science. A core element of future weather (or air quality) forecasts and climate projections is numerical models that not only provide a foretelling of physical indicators of future climate but also indirectly provide information on societal impacts and thus provide a key resource for addressing adaptation and mitigation questions. Because of the critical spatiotemporal data such models offer, it is a high priority to bring as much observational scrutiny to the output from the numerical models as possible. This requires the systematic application of observational datasets from various sources. As such, enabling spatiotemporal analysis of observational datasets and evaluation of the spatial temporal output from numerical models are all necessary for visual analytic systems to provide a reliable characterization of future weather/air quality/climate that can lead to an informed decision-making process.

Data Exploration: There are visual analytic systems whose data processing and exploration leverage various statistical techniques. Piltner et al. primarily focused on spatiotemporal interpolation of irregularly spaced air quality observations for the contiguous US [24]. Their spatiotemporal interpolation and its validation are conducted in the stage of data processing, so users of the visual analytic system explore the interpolated data via a web interface. Similarly, DDLVis [37] applies three advanced statistical techniques (a peak-based kernel density estimation, a dictionary learning method, and a peak-based variation generation model) to store, visualize,

and query climate data efficiently. Climate Engine [43] offers calculation of basic statistics including mean, median, maximum, minimum, and total only. However, this cloud-based web application uses a variety of observations at high spatial and temporal resolutions. Moreover, with special emphasis on the United States (US) National Climate Assessment (NCA), the National Climate Change Viewer (NCCV; [13]) provide various summary statistics for various geographical regions, such as counties, states, and NCA regions, over the contiguous US. The summary statistics include spatially averaged time series and percentile tables of temperature and precipitation from 30 models for present and future climate.

Multivariate Analysis: In most analysis tasks, climate scientists analyze multiple variables which are associated with certain phenomena. For example, concentrations of ozone, nitrous oxides, and particulate matters from observations and numerical models are analyzed in air quality studies. In multi-variate analyses of climate data, it is always important to figure out a normal state, which is usually defined as averages of individual variables over a certain period. Wang et al. [36] applied an association rule-learning algorithm to study the relationship of multiple variables in climate data. To facilitate the rule-learning algorithm and parallel coordinate plotting, they also applied a categorization algorithm to group each variable's values into five categories. Moreover, Poco et al. [42] demonstrated an inter-comparison among simulated climate models leveraging multifaceted data. CrossVis [63] provides a visual analytics system to explore large-scale heterogeneous multivariate data with a use case of historical hurricane observations.

Anomaly Detection: Anomalous events, such as heatwaves, cold surges, heavy precipitation, drought, and severe air pollution, can be detected by analyzing the deviation of spatiotemporal data from their averages. Using the Scalable online visual analytic system (SOVAS; [14]), climate scientists can detect anomalous events such as extreme heat events. SOVAS also enables the calculation of correlation coef-

ficients between two meteorological variables and spatial means. Voila [64] presented a tensor-based analysis algorithm in interactively detecting anomalies in spatiotemporal data. Moreover, the phenomena portal [38] provides a map-based user interface to visualize certain anomalous events detected by their convolutional neural network (CNN) model. Users of the portal can provide feedback on the detected events.

Uncertainty Analysis: The evaluation of numerical models against observations aims to quantify uncertainty for future projections from the models. This uncertainty quantification process is key to informing climate model development and providing actionable climate information to support decision-making [65]. To visualize forecast output from multiple simulations and their uncertainty, Noodles [66] applies two visualization techniques, uncertainty glyphs and an uncertainty ribbon. The authors compared these two techniques with conventional spaghetti plots and show the advantages of the new techniques. Herring et al. [40] presents a context switching technique in ClimateData.US to explore low and high emission scenarios of climate variables such as temperature and precipitation. It essentially enables the uncertainty analysis of climate risks.

Trajectory & Flow Analysis: The trajectory analysis featured in several visual analytic systems is widely used to track hurricanes or air pollutants including volcanic ashes and dust storms. Kim et al. [51] developed a novel flow analysis technique to extract the flow map from the non-directional spatiotemporal data. EnsembleGraph [67] utilizes graph visualization of ensemble simulations. In the graph, nodes are subregions with similar simulation output and edges represent spatial overlap. The Hawaii Rainfall Analysis and Mapping Application (HI-RAMA, [12]) applies a random forest model for quality control of rainfall station data and two different weighting approaches for gap filling of missing data. Again, these statistical techniques are not applied to users' data exploration.

2.3 Visualization Techniques

In this section, we discuss various visualization techniques that have been employed to visualize distributed spatiotemporal data. We investigated these visualization techniques from different aspects to report a categorization and mapping to spatiotemporal data types and analysis tasks discussed respectively in previous sections. We reviewed the taxonomy for visualization techniques defined by Keim [68] and extended further by Afzal et al. [27]. In their work, the visualization techniques are classified as standard 2D/3D displays, geometrically transformed displays, icon-based displays, dense pixel displays, and stacked displays. Spatiotemporal visualizations generally leverage multiple visualization techniques to illustrate the spatiotemporal data due to its high dimensional and multivariate attributes. Hence, display design is also a crucial aspect of visualization for spatiotemporal data.

In our review, we provide a classification of the visualization techniques inspired by [68] albeit based on the display design instead of visual encoding. The classification focuses more on the different approaches of integrated visualization methods for visualizing spatiotemporal data. For the geo-spatiotemporal data, most visualization techniques leverage a layered satellite map to visualize the geospatial portion of the data. Choropleth maps are also widely used for the spatiotemporal data where the spatial portion includes vector data [69, 13]. Moreover, the temporal portion of the data is generally mapped either to a linear timeline representation [70, 3] or a third display dimension [42, 47]. To characterize the visualization techniques for spatiotemporal data, we first discuss the different standard 2D/3D visualizations such as time series plots [38], bar charts, parallel-coordinates [42, 63], volumetric rendering [36, 34], map visualizations [31, 71], etc. Then, we discuss different display designs for integrated spatiotemporal visualizations such as temporal transition, path ensembles [72, 73], color coding [34], space-time cube [74], radial map [69], and coordinated multiple views [44, 37]. These integrated visual representations assist scientists in

performing exploratory visual analysis such as sensemaking [12], anomaly detection [64], trajectory analysis [51], and visual query [37].

2.3.1 Standard Visualization Techniques

Spatiotemporal variables generally contain timestamps and geolocation, so geospatial visualization and time series charts are widely used by scientists to analyze their data [75]. Geospatial visualization provides an interface to inspect and explore spatial variables with geographic location [76, 13, 44] whereas time series charts are essential to perceive the temporal trends of the spatial features. The most commonly used techniques to visualize the spatiotemporal data are a combination of standard charts, such as a line chart, a scatter plot, and their variations, [77, 27] along with a map visualization (i.e., satellite [71], choropleth [29]). These standard visualizations are illustrated in 2D and 3D. The basic idea of 2D visualizations in this context is to present the data variable against time in x-y axes. In contrast, 3D visualization includes another dimension essentially supporting the analysis of trends and patterns in spatiotemporal data with x-y-z axes. However, leveraging standard charts often create occlusion on the visual interface that challenges cognition efficiency [78].

2D visualizations are the most commonly used technique to illustrate spatiotemporal data. This category includes basic charts such as line graphs, scatter plots, bar charts, etc. [29, 66, 69, 43, 64, 38]. Li et al. [14] time series plot to present temporal trend and correlation between temperature and precipitation. Sharma et al. [39] leverage a combination of a line graph and a scatter plot in their work to study the temperature anomalies in ocean and land surfaces.

Other 2D charts besides line and scatter plots are also utilized to visualize the spatiotemporal variables. Parallel coordinates are popularly leveraged to analyze the association in multivariate spatiotemporal data. Poco et al. [42] studied the multimodel distribution of spatiotemporal climate variables in a parallel coordinates visualization. Steed et al. [63] demonstrated interactive parallel coordinates consisting of

categorical and numerical axes to explore large-scale heterogeneous multivariate data. Wang et al. [36] and Li et al. [14] also used a multifaced parallel coordinate plot to illustrate the association of climate variables near a cyclone center. Bar charts are also leveraged to compare the multi-model, seasonal and yearly distribution of spatiotemporal variables in [13, 43, 44]. In addition, Shu et al. [67] introduced a 2D temporal graph that provides spatiotemporal behaviors in ensemble simulation data. Similarly, there are other 2D visualization techniques utilized for exploring spatiotemporal variables such as node-link diagram [3], area plot [23], hierarchical clustering [48], etc.

2.3.2 Integrated Visualization Techniques

The display design for the spatiotemporal visual analytics interfaces is crucial as complex spatiotemporal data often challenges human cognition. Several studies investigated the design factors that impact the human ability to comprehend the illustrated data [79, 77, 70]. In this section, we discuss the integrated visualization techniques employed to explore distributed spatiotemporal data.

Temporal Transition: The transition technique in spatiotemporal visualization mostly consists of a geospatial or contour-based visualization to plot the spatial intensity along with a secondary visualization to stream through the temporal dimension. It provides a dynamic representation as the temporal events appear and disappear on the spatial map controlled by time or user's selection [70, 39]. The spatial dimension is illustrated in a 2D/3D geospatial map often layered with a base map or choropleth. The temporal progression of the data variables is summarized using standard visualization techniques such as line graphs, area charts, scatter plots, or a series of geospatial depictions of the temporal data [23, 38, 44, 37]. Time sliders are also a popular choice to demonstrate the temporal transition on a geospatial visualization for spatiotemporal data [80, 66, 81, 40, 14]. Display transition and animation have also been employed in CORNEA [71] where sliders are implemented to control the

time speed, rotation, and altitude on the 3D globe.

Maskey et al. [38] demonstrated a spatiotemporal depiction using the display transition of geospatial visualization. Pahins et al. [23] designed an integrated visualization to explore large-scale spatiotemporal data using a geospatial view and area chart. The transition view allows the user to perceive the subtle temporal change of events in space. However, the fast progression of events can overwhelm the user to explore the data variables. The user attention is also split between the geospatial and timeline view which might lead to missing the non-salient changes over time. Mayr et al. [70] discussed several studies to compare the animation speed and user interactions for spatiotemporal transition to find cognitive efficiency.

Space-time Cube: Space-time cube is an integrated 3D visualization technique leveraged in visualizing various spatiotemporal variables. To answer the question of where (space) and when (time) from the spatiotemporal data, Kraak [82] presented this technique combining specific visualizations for space and time. In a space-time cube, the cube's bottom represents space whereas the height represents time with temporal cutting operations. A spatial map is generally used to denote the space followed by multiple layers of standard visualizations to illustrate the spatiotemporal variables in different timestamps in the dataset [83, 84, 85]. We have reviewed the use of bar plots, scatter plots, volume rendering, and spatial maps on the temporal axis in space-time cube visualizations.

Schroth et al. [80] conducted a case study on the existing tools for understanding the spatiotemporal climate scenarios where they presented a space-time cube for resolving the multi-dimensional data visualization problem. Eaglin et al. [47] provided a web-based interface illustrating a space-time cube where the temporal axis consists of a volume rendering. The user can interact with the 3D view to visualize a time layer from the volume in a 2D heatmap. Poco et al. [42] demonstrated the use of a space-time cube in the SimilarityExplorer to visualize the complexity of multi-

faceted data variables. Hadlak et al. [45] presented various illustrations of space-time cubes using colored links, pencils, and helix glyphs to visualize temporal patterns in spatiotemporal variables. Visual encodings for the temporal axis are able to include multivariate data in a space-time cube visualization. Rober et al. [86] utilized space-time cube representation while demonstrating in-situ visualization [87] processing with spatiotemporal data. Ferstl et al. [74] presented an approach to visualize time-varying iso-contours in a time-hierarchical clustering using juxtaposition. The stacked iso-contours in the space-time cluster essentially form a volumetric rendering of the weather forecast ensembles.

Space-time cube visualizations are able to encode high-dimensional spatiotemporal data in an integrated display design. Hence, this visualization is utilized to provide an overview of the spatiotemporal events to the users. However, with an increasing number of time steps and variables in the data, the visual occlusion can challenge the user's cognition in conducting the exploratory analysis.

Trajectory Visualization: Trajectory visualization approach generally plots multiple temporal layers into an integrated map where transparency and intensity derive the position of the object or event. This technique encodes the temporal dimension with different variations of colors. To denote the spatial boundary, the color coding technique is often merged with a geospatial visualization [51, 24] or choropleth [29, 88]. Various transformations of this technique are leveraged to visualize the data flow, path ensembles, and trajectory in the spatiotemporal data.

Trajectory visualizations illustrate the temporal movement of spatiotemporal attributes in geographical or abstract space. These visualizations can assist scientists to identify valuable patterns and trends from the spatiotemporal data. Various illustrations are employed to explore the underlying patterns in the moving object/event data [89]. Liu et al. [49] demonstrated trajectory visualizations for ensemble representations and position prediction of the storm path. They employed the ensemble path

visualization to denote the hurricane direction and smoothly interpolating hotspots to denote the storm-strike position. Kim et al. [51] leveraged similar techniques to visually interpret the flow map in spatiotemporal data without trajectory information. Wang [90] published a software suite for meteorological data visualization, MeteoInfo, where the wind trajectory is demonstrated with ensemble path visualization over spatial contour from the station data.

Coordinated Multiple Views: Coordinated multiple views (CMV) is an exploratory visualization technique that enables the user to explore the data using multiple visualizations integrated into a window [91]. CMVs are employed to interactively visualize the complex high dimensional spatiotemporal data [92]. In spatiotemporal visualizations, CMV visualize the spatial and temporal aspect of the data by combining multiple visualization techniques. A geospatial map is often used to visualize the spatial portion and a linear representation is used to visualize the temporal portion of the spatiotemporal data [70]. The goal is to allow the users to identify the facts from complex structured data variables by providing an integrated set of standard visualizations without cluttering the window. Visualizations in CMV are integrated through rich user interaction and view manipulation [91].

Wang et al. [36] demonstrated CMV for conducting association rule-based multivariate analysis of spatiotemporal climate data. They employed a 3D volume rendering and parallel coordinates to identify multivariate correlations among spatiotemporal climate variables. Parallel coordinates are leveraged with geospatial spatial heatmap to compare multifaceted spatiotemporal variables in SimilarityExplorer [42]. Li et al. [69] presented CMV in Vismate to visualize the station-based observation data on climate change. Moreover, Voila is a visual anomaly detection and monitoring system where Cao et al. [64] presented a CMV compatible for visualizing the spatiotemporal variables. Li et al. [37] recently published a visual analytics system, DDLVis, for the real-time visual queries of spatiotemporal data. They designed the

interface with a geospatial map and area chart to navigate through the summarized spatiotemporal data. Based on their interaction with these visualizations the results are presented in a streaming view side-by-side.

CMV often cases leverages the standard visualizations that fall under the visual literacy of the domain users. With rich user interactions and view manipulations, CMV overcomes the shortcomings of standard visualizations in extracting underlying knowledge from the data. However, depending on the visual encoding techniques, the hierarchical number of view manipulations, and user interactions, CMV can be overwhelming for the users. Several analytical reviews [93, 92, 91] reported diversity in visual encoding and consistency in context switching are essential in visualizing spatiotemporal attributes in CMV.

2.3.3 User Interactions

User interaction in the visual analysis approaches is essential to facilitate the user with exploratory functionalities [94]. What user interactions mean for visual analysis differs based on systems, tasks, and users' intents [95]. In visual analytics systems, the user interactions affect the pipeline, transform raw and processed data, and alter the mapping and view [95]. In this section, we discuss different user interaction methods demonstrated in spatiotemporal visual analytics. We categorize the interaction techniques based on the interaction methods taxonomy discussed in [94, 96]. These user interactions are often correlated and overlapped while implementing the visualizations.

Select "Select" interaction enables users to mark or tag specific data based on their point of interest. The selection allows to keep a certain set of data visible on the interface or explore a specific set of related data. Moreover, this interaction method can be essential to remove outliers from the viewport.

In visual analytics systems, select interaction is found as the most common user interaction method. Panoply and NCCV [13, 97] both render visualizations for climate

projections where they offer selecting parameters for altering the time entity. NASA's Giovanni [98] provides interaction with the interface to select variables, time range, and an analytical plot that essentially renders a static visualization. Select interaction is extensively leveraged in the CMVs as it allows the user to select an area of interest to update other visualizations accordingly [69, 48, 44]. McLean et al. [12] employ select interaction on the map markers to display information about rainfall station and volume. Select interaction essentially enables other interaction methods on the visualization such as explore, connect, or abstract [66, 42, 47].

Explore "Explore" interaction provides users the control to include new items in the view, usually by superimposing on the interactive view. In the context of spatiotemporal climate variables, the data is often large in volume, and particularly challenging to display all the details in a single window. Explore interaction is leveraged in such cases to allow the user to interact with a view to include an additional item on the window that provides more detailed information. It enables the user to examine and identify interesting data subsets. As a result, explore interaction method is very useful for the exploration of high-dimensional feature spaces.

Explore is a widely used interaction method for integrated visualizations, especially for the map visualizations to support identifying anomalies [64], uncertainties [66], or patterns [48] of the data. Quinan et al. [88] leveraged explore interaction to enable isocontour features on the ensemble map visualization. Johansson et al. [41] implemented explore interaction in their tool where users can explore the risks for flooding and sea level rise interacting with the climate scenarios for a selected location.

Reconfigure "Reconfigure" interaction enables the user to change the arrangement of visualization by sorting or re-arranging. Reconfigure also helps the user remove the occlusion from the window to get a clear visualization of the data. Moreover, the reconfigure interaction can benefit the user performing multivariate visual analysis. Reordering the axis in parallel coordinates visualization can be an example of the

reconfigure interaction method. This user interaction is utilized to support time-varying large-scale multivariate analysis using parallel coordinates [42, 14]. Steed et al. [63] demonstrated parallel coordinates visualization reconfigured with focus and context range selection for the temporal and numerical axes. Wang et al. [36] leveraged the reconfigure interaction in 3D spatial volume rendering to demonstrate the rule-based multivariate analysis.

2.3.3.1 Encode

"Encode" interaction fundamentally alters the visual representation by changing the number of dimensions, colors, and sizes in visualizations. It helps the user gain a deeper understanding of the properties of data. Encode interaction can be particularly essential in multivariate analysis where the user can map the attributes to different colors and shapes. This interaction allows the user to find insight from the data analysis from a different perspective. Eaglin et al. [47] employed encode interaction that allowed the user to select a 2D temporal slice from the space-time cube. Li et al. [34] allowed the user to colorize the trajectories based on the variable and length of the trajectory. Moreover, interfaces employing parallel coordinates also demonstrate the use of encode interaction [63] as it often changes the color based on the selected range of multifaceted attributes.

Filter "Filter" interaction helps the user to analyze data by applying conditions to the rendering. This interaction allows the user to specify a range or condition to visualize a subset of the data. The filter also works for exploring information from the dataset. In general, searching and querying data are the common use cases of filter interaction as it excludes the data from the view that does not satisfy the conditions.

We reviewed a handful of visual analytics systems employed in spatiotemporal analysis that supports interaction to filter the sample space (Task requirement **R4**). DDLVis [37] employed filter interaction for the user to draw an area of interest on the map visualization. Li et al. [14] allowed the user to write a textual query to

visualize spatial heatmap and temporal trends. Moreover, spatiotemporal data such as in climate projections produces complex and high dimensional data that often create occlusion in the view. This interaction is utilized in those scenarios to allow the user to focus on the data of interest [29, 64, 38]. Sanyal et al. [66] allowed the user to filter the glyphs and ribbons presenting ensemble uncertainty on the geospatial visualization. Climate Data Analysis Tool (CDAT) provides scope for user interaction to filter the data selection range that essentially reflects on the visualization [99].

Abstract "Abstract" (or elaborate) interaction allows the user to view data from various levels of granularity. This method is useful for working with large datasets, as it can provide an overview of different stages according to the user's preference. It also provides an explanation of why the sample belongs to a particular cluster. User interaction with the data points to see details on a tooltip and zooming over the map are the common use cases of abstract interaction. Huntington et al. [43] employed abstract interaction in Climate Engine to allow the user to zoom on the time series and view values at the data points. Castruccio et al. [71] utilized abstract interaction in a portable virtual reality environment to explore the spatiotemporal variables across the globe and selected surfaces. [39]

Connect "Connect" interaction enables the user to find relationships between the views and highlight features that are similar or relevant. This interaction is applicable for both single-view and multiple views visualizations. In a single view, connect interaction highlights the related nodes or data points within the visualization. In multiple views, the user interacts with a record to effectively highlight all the related records across multiple views. Spatiotemporal visual analytics interface generally consists of multiple views [34, 38] as spatiotemporal data is often split into spatial and temporal dimensions to separate plots. VisAdapt [41] presents connect interaction where users interact with the geospatial contour to visualize the area detail on the other coordinated views. Wang et al. [36] demonstrate connect interaction where the

spatial volume rendering updates based on the user's interaction with the multivariate parallel coordinate plot.

Shepherd "Shepherd" interaction is the final interaction method that allows users to guide the modeling process. Such guidance can be direct or indirect. Direct guiding enables the user to set or update different parameters for the modeling whereas Indirect guiding includes providing constraints and thresholds. This interaction is not a common method in spatiotemporal visual analytics. Kappe et al. [48] demonstrated decadal climate prediction where they allowed the user to toggle the parameters to refine the ground truth in order to get the most accurate prediction. Cao et al. [64] enables the user to provide context-guided input for ranking anomalous patterns in streaming spatiotemporal data. Li et al. [37] allowed the users to guide real-time queries in spatiotemporal data distributions using density dictionary learning.

2.4 Visual Analytic Systems

Visual analytics include data preparation, pre-processing, analytical workflow, interactive visualization, and usability metrics of exploratory analysis. In this section, we review the visual analytics approaches from the aspect of tools that are leveraged in the implementation process, system accessibility, and exploratory analysis workflow.

Cashman et al. [100] presented a user-based visual analytics workflow for exploratory analysis that enables the users to discover underlying knowledge from the given dataset. The workflow consists of several fundamental steps such as data interaction, problem exploration, generating problem specification, and model exploration. Cui [101] provided a comprehensive analytical review on the visual analytics systems that summarizes the analytics process in 6 steps - data pre-processing, analysis method, visualization, knowledge generation, interactive hypothesis building, and visual reflection of users' perception. Sacha et al. [102] illustrated a knowledge generation model for visual analytics that represents both computer and human-centric

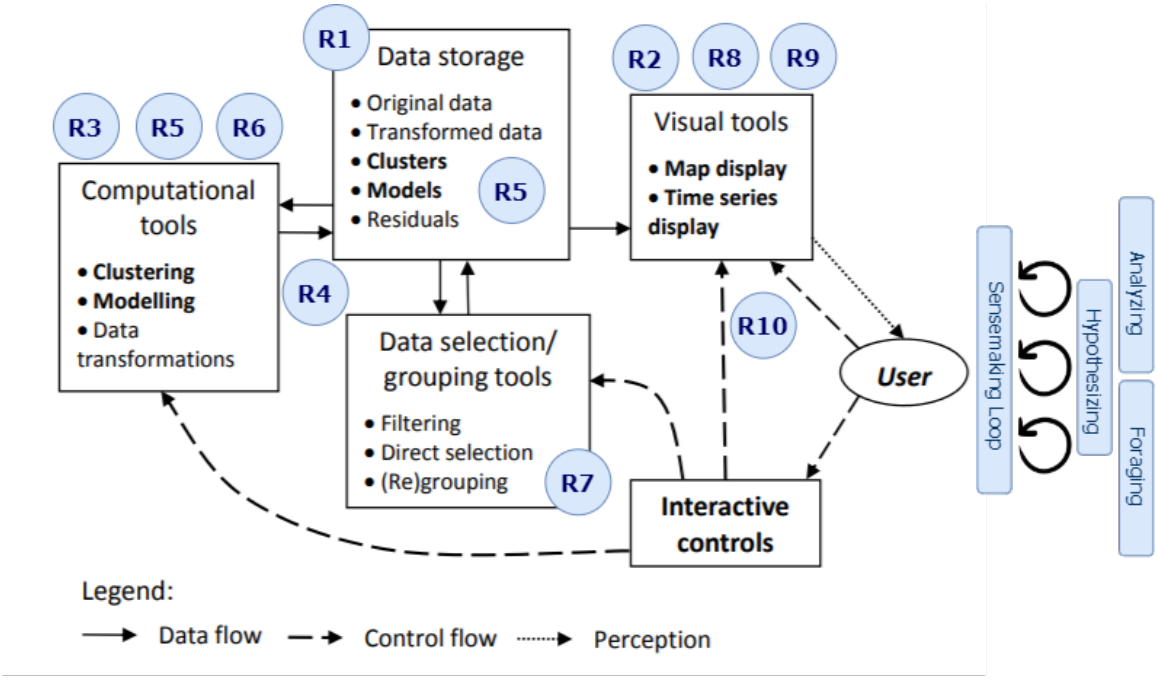


Figure 2.2: Spatiotemporal visual analytics workflow for exploratory analysis [4]. We contextualized the workflow for climate science by labeling it with our identified task requirements for exploratory spatiotemporal visual analytics.

approaches through exploration, verification, and sensemaking loop. Andrienko et al. [4] described a visual analytics framework for spatiotemporal analysis and modeling. In addition to the conventional exploratory visual analytics workflow, this research addressed the user's perception of data and control flow from a spatiotemporal analysis and visualization perspective. Based on our extensive literature review, we believe Andrienko's visual analytics illustration [4] more accurately outlined the visual analytics workflow for spatiotemporal exploratory analysis in climate science.

While reviewing the implementation process, we found that visual analytic systems often leverage other data analytic tools to run the exploratory analysis. These tools mostly enable scientists and researchers to run analytical scripts and visualize the result with basic visualizations. It is popular among scientists to leverage tools such as MATLAB [103], GrADS [104], or NCAR Command Language (NCL) [105]. Moreover, exploratory analyses for spatiotemporal data frequently leverage the capabilities of distributed computing [13, 51, 64, 106], to overcome the complex

and time-consuming analysis. Several systems such as UV-CDAT [28], DataONE [107, 108], SciServer [109], and Open Science Grid [110] are widely used by the analyst and research community that provides shared computing resources to run analyses in the spatiotemporal data. These analysis tools often lack interactive exploratory visualizations and are unable to satisfy use-case-specific visualization requirements.

In contrast, visual analytics systems enable scientists to perform their analysis tasks and explore complex spatiotemporal data with the support of an interactive interface. We label the spatiotemporal visual analytics systems into two categories - general purpose and case-specific. General-purpose visual analytics systems support analysis for numerous use cases and data set. These systems generally demand the dataset according to a pre-defined structure. Climate Engine [43], SOVAS [14], and WebGlobe [39] are some of the leading examples of general-purpose visual analytics systems. However, these systems often fall short of satisfying specific analytics or visualization requirements for case-specific visual analytics systems. Scientists work with heterogeneous data that frequently demand custom features to analyze and explore the underlying trends and patterns. Most special purpose systems have been developed with the involvement of the domain experts injecting their specific requirements [111]. In our review, we also observed domain researchers collaborating with visualization researchers to develop visual analytics systems that support case-specific exploratory analysis of spatiotemporal data. Noodles [66], VisAdapt [41], DDLVis [37] demonstrate examples of case-specific visual analytics systems. While these systems demonstrate components and functionalities for case-specific analysis, the analytics interface and system infrastructure share the common challenges of accommodating distributed large-scale spatiotemporal data sets. There are plenty of research gaps to be addressed by visualization researchers in this context.

2.5 Discussion

In this dissertation, we organized the literature identifying key aspects of exploratory spatiotemporal visual analytics systems using our diverse domain knowledge from experts and understanding gained from the literature review [68, 112]. Six dimensions were identified. **Analysis Techniques** - Where scientists propose novel analysis techniques to explore the potential impacts of climate change [113], health fairness [114], urban planning [15], etc. The proposed techniques use visual analyses to study spatiotemporal variables such as precipitation, temperature, and air quality [115, 19]. **Data Source & Definition** - Where data scientists generate high resolution geospatial projections [17, 81]. We organize this dimensional according to data structures and formats utilized in the context of climate science. [52, 59]. **Analytic Systems** - Where novel analytics pipelines are presented for visual analytics systems (e.g. [116, 24]) to identify key features and limitations in supporting data analysis for the exploration of spatiotemporal data. **Visualization Approach** - Where novel exploratory visualization approaches [72, 117, 46] are presented to create a map and compare it against corresponding analysis tasks. **Interaction Approach** - Where interactive features facilitate the scientists' efforts to explore, reconfigure, and compare the visual renderings of spatiotemporal data. This dimension emphasizes the approach's potential to unfold knowledge from multivariate spatiotemporal attributes, e.g. [118, 3]. **Evaluation** - Where the challenge of defining quantitative and qualitative metrics to evaluate visual analytics systems is the focus [119]. We organize this dimension according to performance metric, analysis task, and target user group [120, 70] to close the gap between usability and task efficiency.

CHAPTER 3: DISTRIBUTED VISUAL ANALYTICS FRAMEWORK

In this chapter, we present a visual analytics framework that addresses the complex user interactions required through a command-line interface to run analyses in distributed data analysis systems. The visual analytics framework facilitates the user to manage access to the distributed servers, incorporate data from the source, run data-driven analysis, monitor the progress, and explore the result using interactive visualizations. We provide a UI embedded with generalized functionalities and access protocols and integrate it with a distributed analysis system. To demonstrate our proof of concept, we present two use cases from the earth science and Sustainable Human Building Ecosystem research domain.

3.1 Introduction

In support of sensemaking, users require a visual analytic interface that seamlessly supports data discovery, exploration, and analysis. In other words, the visual analytic interface should support the full extent of the sensemaking loop [121, 122] from foraging to hypothesizing to analyzing. Current solutions, however, often emphasize specific aspects of sensemaking â for example, data exploration or data analyses - and fail to support the full analytical lifecycle adequately. In addition, it is infeasible to access large and remote datasets using traditional pipelines for data transformation, conversion, and presentation. Such pipelines are commonly preceded by massive data downloads, which are infeasible or impractical for many remote datasets. Thus, new pipelines are required, pipelines that are not predicated on massive data downloads. Finally, to generalize the visual analytic interface for distributed fragmented data, new APIs and data access protocols are necessary. In particular, these APIs and protocols must account for the full analytical lifecycle and must not be predicated on massive, upfront data downloads.

In this chapter, we present an interactive VAF for distributed data analysis systems

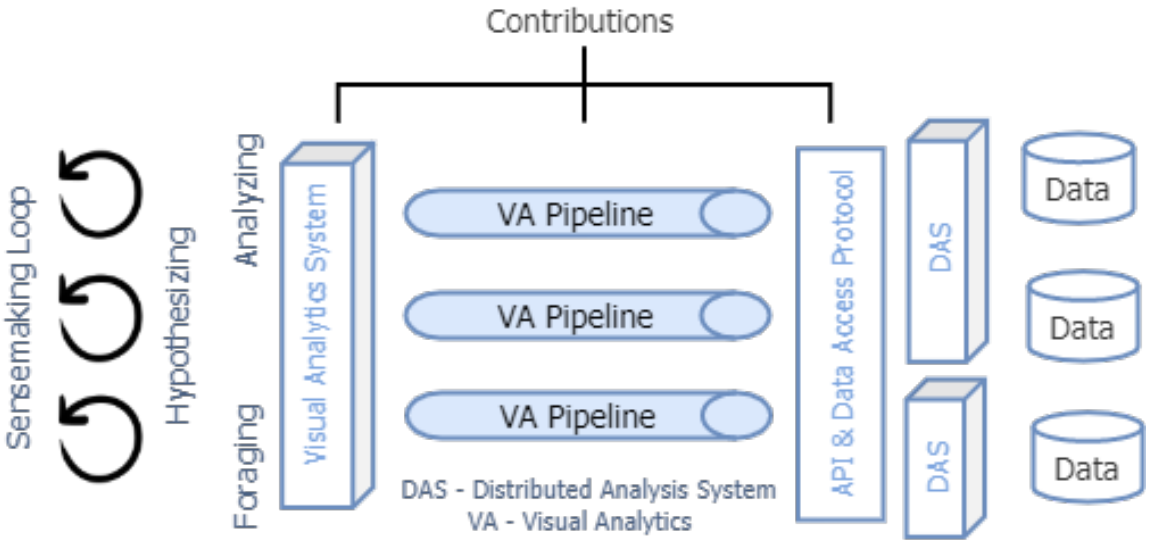


Figure 3.1: Visual analytic system pipelines for distributed analysis systems.

(DAS). VAF enables analyses over distributed, fragmented data without the movement of massive data. Significant advancements in distributed data analysis over the past decade [123, 109, 124, 125, 126] make our proposed framework a feasible candidate to accelerate the analysis tasks of researchers and analysts. To demonstrate our framework, we leveraged the Virtual Information-Fabric Infrastructure (VIFI) [123, 127, 128, 129, 130, 131, 132], which is a computational infrastructure that enables analyses across distributed, fragmented data without the movement of massive data. Within VIFI, analyses migrate to the distributed data, and only derived data - e.g., result sets - migrate from the data hosts. The main contributions in this chapter are:

- We define a VAF for distributed, fragmented data as well as design goals and associated implementation tasks.
- We present a generalized pipeline for data transformation, conversion, and presentation - one that is not predicated on massive, upfront data downloads.
- We provide a demonstration version of the visual analytic UI to support distributed analysis.

- We present generalized APIs and data access protocols to enable proper integration with infrastructures that enable analytics over distributed fragmented data.

- We demonstrate VAF with two analytic systems (i.e., VIFI and simple file-based systems) and illustrate its benefits using two use cases from earth science and Sustainable Human Building Ecosystem (SHBE) research domains.

3.2 Related Work

There is an abundance of previous research (e.g., [107, 108, 110, 109, 133, 134, 135, 123]) spanning many disciplines that demonstrates the potential value and impact of enabling analyses and sensemaking across distributed, complex, and fragmented data. Yet, significant challenges remain. In particular, to support sensemaking across such data, new visual analytic interfaces are needed, new pipelines for optimized distributed data interaction and visualization are required, and new data access protocols and application programmer interfaces (APIs) must be developed. We highlight these challenges in Figure 3.1.

Current data-driven applications often require the identification and mitigation of relevant data from multiple locations to a common storage location, prior to performing analysis. To overcome what is often a difficult, time-consuming, and laborious task, some alternate solutions have been proposed for data sharing using high-speed networks and cloud-based hosting, while other alternative solutions focus on providing shared computing resources. DataONE [107, 108] is a project focused on providing easier access, search and discovery to earth and environmental science data repositories. The Open Science Grid [110, 136] enables scientific research by providing distributed computing resources. SciServer [109, 137] is a cyber-infrastructure system that provides a suite of tools and services (including storage, access, query, and processing) for big data analyses from various disciplines leveraging data with dif-

ferent format and structure. While SciServer collects all data at a common storage location, it attempts to minimize data movement by collecting data at the location that contains the majority of the required data. SciServer also migrates the analyses by sending Jupyter Notebook [138] to the common storage location.

Other data-driven applications aim to develop research infrastructures that integrate storage, high-performance computing, and analytic tools (e.g., XSEDE [124, 139], NeCTAR [140], PRACE [141], and EGI [142]). The applications allow end-users to share distributed computing resources and data repositories. The solutions may be used by Science Gateways (SGs) [143, 133, 144, 145, 146] to provide (web) portals and UIs that enable scientists (e.g., chemists, biologists) to access, build and execute analytic workflows. SGs relieve scientists of the burden and needed the expertise to set up and maintain the underlying distributed cyberinfrastructure. SG services can be shared and reused by different end-users. SGs can be classified into SG frameworks like WS-PGRADE/gUSE [143], and SG instances like the computational neuroscience gateway [126]. SG frameworks are generic SGs that provide low-level services for scientists from different domains. While SG frameworks provide high-level abstractions for computing specialists, SG frameworks require additional learning from scientists to leverage their full potential of the frameworks. SG instances provide high-level services for scientists in a specific domain. Thus, SG instances simplify scientific operations for end-users but limit flexibility when more functionalities are needed from the SG instance. Some of the SG features and services (e.g., security, data and workflow management) depend on the underlying technology. Thus, it becomes challenging to port an SG from one infrastructure to another [147, 148]. Gugnani et al. [149] suggest a generic approach to integrate infrastructure-aware workflows, (e.g., WS-PGRAD/gUSE [143]) with bigdata parallel processing tools (e.g., Hadoop). This work [149] uses the CloudBroker platform [150] to provide required cloud-based computational resources.

SGs can be accessed through different middleware like Airavata [151], Agave [152], and Globus [153, 154, 155]. Airavata [151] allows users to manage applications and workflows on the provided resources (e.g., clouds, cluster, grids) through component abstraction of major tasks. The system components are indirectly accessed through component APIs. Agave [152] provides web access, through Representational State Transfer (RESTful) APIs [156], to given resources (e.g., HPC, cloud) to run analyses and to manage data. Globus [153, 154, 155] is software-as-a-service designed to make it easier to discover, replicate, and access big data resources at different locations. Globus is used to deliver scalable research data management services in a secure manner to a variety of stakeholders. Some Globus features, like data publication and managed endpoints, include licensing fees.

In contrast to existing solutions, our VAF aims to support "truly distributed analytics" where analytics are executed at data sites without the massive movement of data. Our framework avoids huge data transfer times while complying with owner-defined authentication and authorization policies for data access. Our framework does not add new infrastructure for additional data and/or computational operations; rather, it aims to integrate with existing data site infrastructure. The framework utilizes containerization technology (e.g., Docker [157, 158, 159, 160]), rather than tools like Jupyter notebooks [138], to migrate analyses. This provides more flexibility over the analytics tools and analytic environments that can be used by the scientists in conducting data-driven inquiries (i.e., analyses are not limited to the tools provided by Jupyter). In addition, unlike some related work, our framework depends entirely on open source technology. For example, our pipeline uses only open-source components (e.g., Apache NiFi [161] and Docker Swarm [162]) with free access to all features. Thus, users can develop, reuse, and customize our framework for their needs.

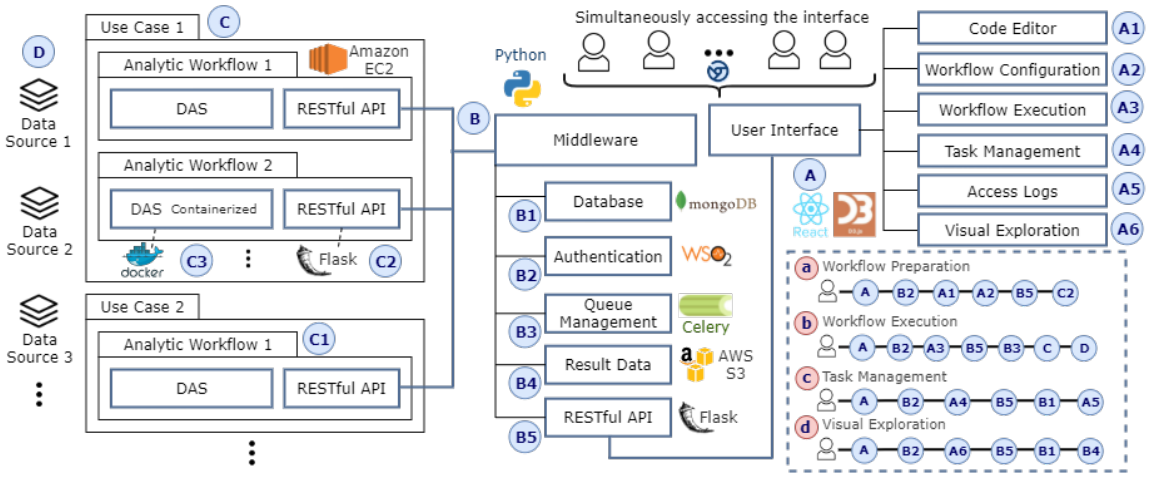


Figure 3.2: Proposed VAF pipeline for DAS: A) UI, B) Middleware, C) DAS site, and D) Data site are the main modules in our framework. a) Workflow Preparation, b) Workflow Execution, c) Task Management, and d) Visual Exploration show the flow of interactions within the VAF components.

3.3 System Design

This chapter illustrates an interactive VAF to simplify user interactions and enhance the user experience with a DAS. To design a pipeline for the VAF, we reviewed numerous distributed analytic systems (e.g., [123, 109, 99, 163]) to identify the key user interactions required to operate these systems. We discovered that many systems utilized command-line interfaces. Nonetheless, we extracted the following fundamental interactions: managing access to distributed servers, preparing analytic scripts and runtime environments, importing data from remote sources, executing analyses, monitoring the execution progress, and inspecting and exploring the analytical results. DAS commonly maintains data-site to data-site communication using cloud infrastructures to run analyses [123, 99, 127]. Operating a DAS from a command line interface requires access for a user to multiple remote servers. Access control for such interaction with the data sites and DAS sites can be complex for the data owners. Consequently, the entire procedure to run a data analysis can be similarly challenging for the data analysts and the end users. Moreover, to explore the results, users from different domain areas were required to pull the resulting data from the

server. Rather than using command line interfaces, DAS often provides a visualization toolkit [99, 164]. However, users are responsible for generating exploratory visualizations or necessary artifacts to measure the performance of the analysis [165]. Given all of these need interactions and associate limitations of current solutions, we identified associate design requirements and implementation tasks and mitigate current complexities for user-DAS interaction.

3.3.1 Design Requirements

In this section, we outline the requirements for an interactive VAF to provide more seamless user interaction with distributed analysis systems. Related work reveals the following design requirements for our VAF:

DR1 **To mediate user interaction with distributed servers.** The framework should provide sufficient features to allow users to execute analyses in DAS without requiring direct user access to the distributed servers and data hosts.

DR2 **To provide a unified model for authentication and access control for distributed servers.** The framework should provide proper access to data and analytic workflows according to data site policies. The framework should integrate with existing authentication and authorization mechanism to the computing servers and various data sites.

DR3 **To enable the exploration of data and resulting analyses using interactive visualizations** The framework should utilize interactive visualizations to support the sensemaking loop (i.e., foraging, hypothesizing, and analyzing) while not requiring massive data downloads as a means to enable accurate and reliable revelation of latent, complex correlations, patterns, relationships, and such other knowledge.

3.3.2 Implementation Requirements

To address the above design requirements, we identify the following implementation requirements for our framework:

R1 **To provide an interface to manage analytical scripts and Portable Analytic Containers (PACs).** Framework users must be able to access, specify, and manage analytical scripts that are stored in an external repository â e.g., at a DAS data host. As such, the framework should offer an end-to-end synchronization with the available analytic scripts and PACs in DAS **(DR1)**.

R2 **To enable user efforts to configure analytical scripts and workflows.** To conduct analysis across distributed, fragmented data, coordinated execution of analytical scripts is often required (hypothesizing). Workflows often contain a set of configurations that points to the dataset, analysis scripts, required access credentials, etc. The framework should provide affordances for users to modify analytical workflow configurations **(DR1)**.

R3 **To support user-initiated execution of analytical workflows in DAS.** After enabling the preparation analysis scripts and configuring an analytical workflow, the framework should allow the user to initiate workflow execution. In addition, the framework should minimize the need for the user to authenticate directly to each data host (e.g., mediate authentication via single sign-on) **(DR1, DR2)**.

R4 **To mediate and comply with data host authentication requirements and authorization policies for datasets, analysis scripts, and workflows.** The framework should manage compliance with authentication requirements and authorization policies for end users. Users should be able to view, modify, and execute analysis scripts and workflows on permitted datasets according to data host authorization policies **(DR2)**.

R5 **To maintain user awareness of workflow execution status.** Workflows often require significant time to queue and execute. The framework should maintain user awareness of workflow execution status so that users may accurately track their progression in the DAS **(DR1)**.

R6 **To provide access to the runtime and error logs.** Runtime logs are useful for the users to understand DAS performance and anticipate expected runtimes of analytical workflows. Similarly, error logs are helpful to trace script and workflow execution, particularly in exceptional circumstances. The framework should effectively present runtime and error logs to users **(DR2, DR3)**.

R7 **To provide an interactive visual analytic interface to support data discovery and explore analytical results.** The framework should provide users with interactive visualizations to discover data (foraging) and explore workflow results (analyzing). The visualizations may be general-purpose or analysis-specific. Thus, the framework should be extensible to accommodate analysis-specific visualizations **(DR3)**.

To satisfy the design and implementation requirements for the proposed VAF for DAS, we developed: interactive, web-based, visual analytic interfaces; a visual analytic pipeline; and, an API / data access protocol.

3.4 Visual Analytic Framework

In this section, we present the three main components of the proposed visual analytic framework. The components include a middleware service, a visual analytic interface, and a distributed analysis system where we utilized both file-based and cloud-based information fabric infrastructure.

3.4.1 Middleware

The middleware for the VAF is one of two major components of the visual analytics pipeline as well as the implementer of the data access API and protocol (Figure 5.3B).

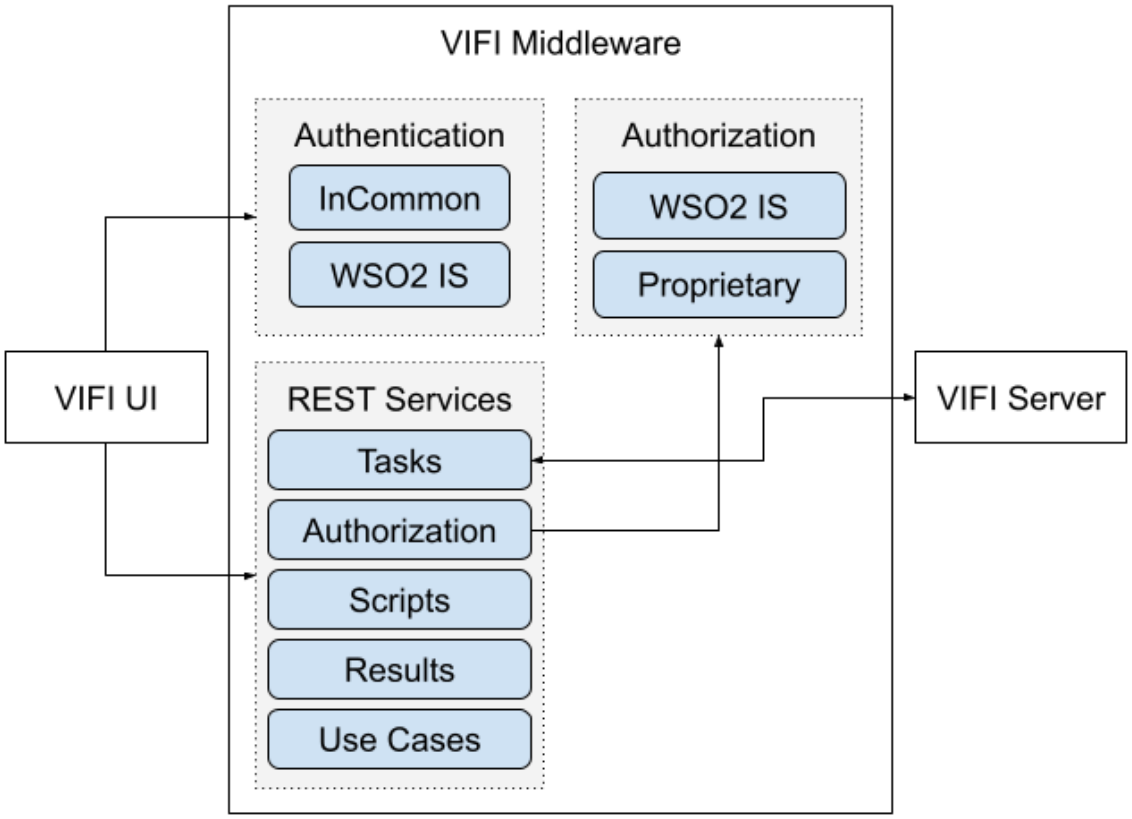


Figure 3.3: DAS Middleware Architecture

It orchestrates use case **(R1)**, workflow/task **(R1-3, R5-6)**, and script **(DR1, DR2)** management as well as authentication **(R4)** and authorization integration **(R4)** with DAS. The primary component is the Task Manager that mediates communications between the VAF and DAS. Figure 3.3 summarizes the primary functions of the VAF middleware. Each is discussed in greater detail in the following.

Use Case Management: Use case management provides methods for creating, modifying, and projecting use cases. A use case organizes a collection of analytical workflows and results. The Task Manager creates a unique key for each new use case. The user, then, specifies a name and one or more workflows (Figure 5.3B1). Results from executed workflows are also collected in a use case. As such, use cases provide a means to organize analyses.

Workflow/Task Management: Workflow/task management provides methods

for the creation, mutation, execution, and projection of workflow specifications and execution instances. Executing workflows is called tasks. Each task is attributed with script identifiers, user identifiers, use cases, and workflow. When a user submits a request to execute a workflow, a new task is created and scheduled for execution via the DAS **(R2, R3)**. The Task Manager collects the required information and relates the information to a unique identifier corresponding to its workflow and use case, respectively. The task along with its related scripts are, then, sent to the DAS for execution (Figure 5.3a). Task information, including execution steps and status updates, is captured in the runtime and error logs **(R5, R6)**. Once a task completes, the Task Manager retrieves the analytical results from the DAS.

Script Management: Script management provides methods for the creation, mutation, and projection of scripts. Scripts and their related configurations are associated with each workflow/task. The script identifier is used during the task creation process to ensure all relevant analyses are properly identified and subjected to the DAS for execution **(R1)**. The analytics interface leverages use case, workflow/task, and script management collectively in the Task Manager to support hypothesizing activities as part of the sensemaking loop.

Results Management: Result management provides methods for the projection of analytical results (e.g., task results). Results for each workflow are associated with a task identifier. When the execution of a workflow completes, the DAS signals the completion status to the Task Manager **(R5)**. The Task Manager, then, retrieves results from the DAS so that these may be projected to the user via the visual analytics interface **(R7)**.

Authentication: To meet the VAF authentication requirements, the middleware uses InCommon [166] and WSO2 [167] for identity management. The VAF, leveraging these services, implements key-based authentication to enable trusted communication between VAF and DAS components **(R4)**.

InCommon is a federated identity management service provided to education and research institutions using the Shibboleth single sign-on architecture. Given a large number of participating institutions and simplicity of setup, VAF integrates with InCommon-based authentication services [166].

For users whose institution is not a member of the InCommon federation, the WSO2 Identity Server (IS) is utilized for authentication. The WSO2 IS integrates with any IAM-compliant architecture. For users with no IAM-compliant architecture, WSO IS provides a built-in IAM architecture. While WSO2 integrates with Shibboleth SSO and thus may be integrated with InCommon, the current VAF implementation leverages InCommon outside of WSO2 IS to simplify configuration [167]. For VAF configurations that leverage WSO2 IS, the middleware authorization service uses the WSO2 API to handle user authorization requests. In such implementations, user authorizations are configured using the WSO2 IS Administration application.

Within VAF, key-based authentication enables trusted communication among VAF components and between VAF and DAS. For WSO2 implementations of VAF, key-based authentication also is enabled between middleware services and WSO2 services. Key-based authentication leverages Hypertext Transfer Protocol Secure (HTTPS) and requires valid certificates for communication between endpoints.

Authorization: The middleware provides two options for authorization support, either a proprietary solution or a WSO2 implementation. For VAF configurations that leverage WSO2, a proprietary authorization solution is provided via a middleware authorization service. Setting up user authorizations using the service requires manual database updates.

We defined three user roles for VAF: a data owner, workflow designer, and data analyst **(R4)**. A data owner manages the user's access control in DAS data sites. A Workflow designer set up an initial configuration and orchestration path for a new workflow. While a data analyst is authorized to tweak certain configurations

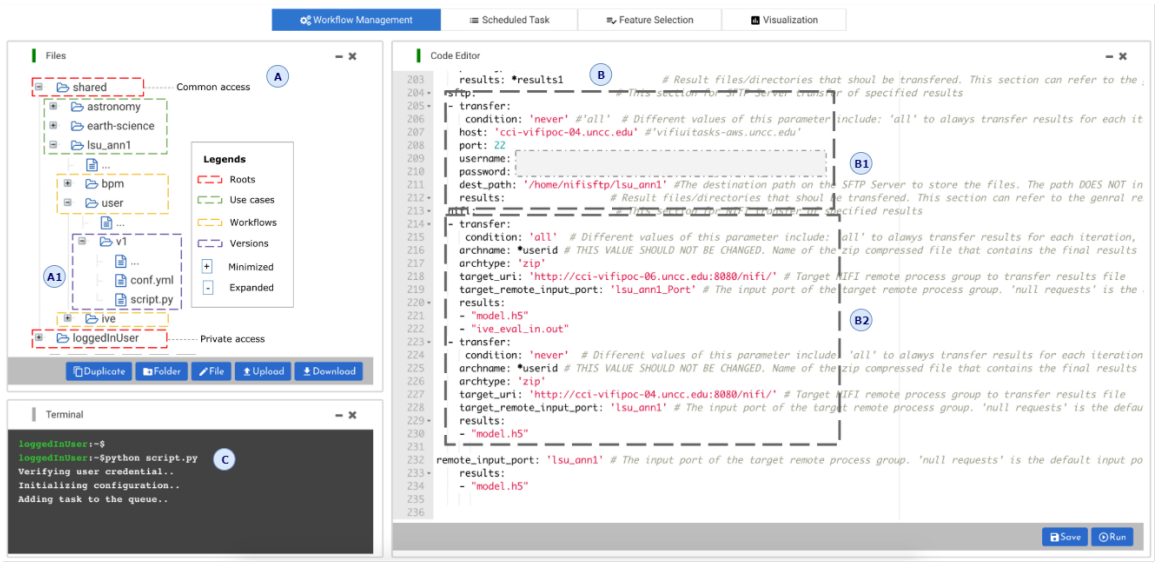


Figure 3.4: Analytic workflow management through our visual analytics systems. Workflow management view provides A) File Browser - to configure the workflow, B) Code Editor - to prepare and run the analytic workflows, and C) Terminal - to stream raw outputs.

according to need, the designer's role is to use the workflows to conduct analyses.

3.4.2 Visual Analytic Interface

The visual analytics interface is the second major component of the visual analytic pipeline. It provides coordinated views [92] to support user actions for workflow execution and result exploration in DAS. To satisfy the design requirements, the UI introduces three main panels: workflow management, task management, and result exploration. These panels assist users in three different phases of sensemaking: 1) data exploration (foraging); 2) analytical workflow and script development and execution (hypothesizing); and, 3), exploring and analyzing workflow results (analyzing). In the following sections, we illustrate support for each phase by presenting VAF support for workflow/script management (hypothesizing), task management (hypothesizing), and interactive visual exploration (foraging and analyzing).

3.4.2.1 Workflow/Script Management

The workflow/script management view consists of a File Browser, a Code Editor, and a Terminal View (Figure 3.4A,B and C). In the File Browser, the available use cases and workflows are listed according to user access privileges to the PAC repository **(R1, R4)**. By default, the view provides access to two types of directories: shared directories and user directories. The shared directories contain all use cases and workflows that are shared with other users. The user directories contain the use cases and workflows (created or cloned) that are private to the user. The File Browser is synchronized with the middleware's Task Manager component via RESTful API. The workflow/script management view presents only those use cases and workflows that are configured in the DAS and flagged as enabled in the Task Manager. We require hierarchical presentation of PACs in the associated DAS as shown in Figure 3.4A. The hierarchy is set in a manner that always gives an ordered path (`/[ root-directory ]/[ use-case ]/[ workflow ]/[ w-version ]/`) when users select a workflow to execute. For example, if the user decides to execute the version 1 of the user workflow shown in Figure 3.4A, the conceptual path to the script directory would be `/shared/lsu_ann1/user/v1/`. The hierarchical abstract organization is adopted for its familiarity and ease of use. Moreover, it provides an encoding that facilitates interface middleware communications.

We added operations to the File Browser (Figure 3.4A) to create, duplicate, or modify the workflows **(R1)**. To keep the integrity of the file structure, each operation is implemented with a set of constraints (Figure 5.3a). The "Duplicate" operation allows the user to clone a selected workflow. It also allows users to clone scripts. For example, in Figure 3.4A1, `ive2.py` is duplicated (or cloned) from `ive1.py`. However, this operation does not allow users to clone use cases or the root directory. Similarly, "Add folder" only allows users to create new version folders under a selected workflow, rather than creating a folder at an arbitrary location in the hierarchy. The "Upload"

and "Download" actions allow the user to migrate analysis to and from the local machine and the DAS.

In the Code Editor (Figure 3.4B), the user can modify the workflow, and create and modify scripts according to their hypotheses for the corresponding use case. By selecting a script, users are allowed to modify and execute the script within the Code Editor for testing purposes (Figure 3.4B). The File Browser also provides access to workflow configurations, which users can select to modify in the Code Editor **(R2)**. In the File Browser (Figure 3.4A), the scripts and workflow configurations are validated prior to execution to assess whether modifications are permitted. The `conf.yml` file associated with each workflow version contains the workflow specification and identifies the appropriate DAS for execution. This file includes, among other things, the DAS credentials (Figure 3.4B1), dataset identifiers, and the location where workflow results are to be transferred after task execution completes (Figure 3.4B2). The Terminal (Figure 3.4C) reflects the output from an associated command line interface to the DAS (when such an interface exists). It also shows log files and the output of test script executions.

To execute a workflow in a DAS, the user selects the `conf.yml` file for the workflow in the File Browser (shown in Figure 3.4A1) and clicks the "Run" button located at bottom right in the Code Editor (Figure 3.4B) **(R3)**. The interface, then, passes the command to the middleware and switches to the Task Management view once execution is launched in the DAS (Figure 5.3b).

3.4.2.2 Task Management

The task management view contains a Scheduled Task panel that lists the workflows (i.e., tasks) that are currently executing for the given user as shown in Figure 3.5. The Scheduled Task panel provides graphical indicators of task progression. A unique task ID is generated for each workflow execution **(R5)**. While executing the workflow, the task identifier is linked to all runtime data, including the runtime environment,

#	TASK ID	Use Case	SCRIPT	TYPE	SCHEDULED TIME	STATUS	LOGS	PROGRESS	ACTION
1	5c530a648423d034715ba2	earth-science	dcp-summary/pr_seasonal_summary.py	python3.6	2021-06-10 15:12:58				Result Run
2	5c548aa9823d0336b1d14	astronomy	CRTS/transfer_learning_latest.py	python3.6	2021-06-10 15:36:03				Result Run
3	5c570ee6823f036ea345b	earth science	dcp-contour/generate_geojson.py	python3.6	2021-06-09 18:09:13				Result Run
4	5c571641823d0374741ad	lsu_ann1	ive/ive1.py	python3.6	2021-06-08 14:33:41				Result Run

Figure 3.5: The UI for submitted analysis task management in our visual analytics system. Task management interface allows the user to interact with the scheduled tasks for inspecting logs, tracking progress, visual exploration of the results or re-running the workflow.

script directory, logs, results files, etc **(R6)**.

A task may take anywhere from fractions of a second to hours or days to execute depending on the size of the data, the complexity of the analysis, the computational resources available, and the shared demand for the data and computing resources. While a task is executing, the user can interact with any tasks to inspect execution logs or view the results of completed tasks (Figure 5.3c). The execution logs accessible from the Task Manager are not the output logs from the given script. Rather, these logs, retrieved from the middleware, capture workflow progression checkpoints for a given task, such as: a) queued â execution request sent to middleware; b) queuing - middleware retrieving relevant scripts, preparing for task execution, generating the unique task identifier, etc.; c) created â the workflow execution request has validated the request and the task is properly created; d) sending - transferring the task to the appropriate DAS; e) sent - the task is successfully sent to the DAS and awaiting execution, and f) complete - the DAS completed the task execution and results are returned to the middleware for user access.

The progress bar aligned with each task in the table (Figure 3.5) depicts an estimation of overall execution progression. The Scheduled Task panel provides users with several operations that may be applied to a given task, including a) ‘Cancel’ â this operation allows a user to cancel task execution by the DAS, b) ‘Rerun’ â this operation allows a user to rerun a task, possibly with updated parameters, after first

Figure 3.6: The variable exploration panel on the UI for analysis results. A) Metadata, B) Triangle matrix-variable correlation, and C) Data variable properties to familiarize the user with the results.

canceling the current execution; and, c) 'Result' â this operation, available after task completion, takes a user to interactive visual interfaces to explore the data that result from task execution.

3.4.2.3 Visual Exploration

The interactive, visual exploration views provide a threefold means to explore both data/datasets (foraging) and task results (analyzing). In this section, without loss of generality, we focus our presentation on results exploration (Figure 5.3d). The interactive, visual exploration views include two principal panels: the variable exploration panel and the visual exploration panel **(R7)**.

The variable exploration panel provides a view that allows users to explore the properties of resulted data. Figure 3.6 shows a sample illustration of the variable exploration panel using this data [168]. The data variable exploration panel initially provides the data dimension (Figure 3.6A), a triangle matrix (Figure 3.6B), and a

data table containing the variable properties (Figure 3.6E). We implemented this panel recognizing that users may not always be familiar with the data variables. This panel provides the data type for each variable in the data. In addition, for numeric data variables, the table provides some statistical data (e.g., range, mean, and standard deviation), though this may not always be relevant or useful. For categorical data, the panel provides count and frequency information. For example, hovering over categorical data presents a bar chart providing the frequency distribution of the categorical data. Additionally, the matrix (Figure 3.6B) provides the correlation among data variables, which may help users during analyses. The matrix cells are color-coded and denote the correlation -1 to +1 using a red-yellow-green color scheme. The user can explore the correlation between two variables by hovering the mouse over the corresponding cell in the triangle matrix. The scatter plot and bar chart (Figure 3.6C, D) based on the respective interactions with variable properties (Figure 3.6B1, E1) allow users to identify and explore patterns or outliers in the data.

The data transformation capabilities include scaling the data variables, applying statistical summary or formula to transform data variables, and injecting domain knowledge to nudge the exploration panel in identifying relevant visualizations. Additionally, the UI allows the user to input thresholds such as good, moderate, and poor correlations, standard deviations, and minimum and maximum factors for the unique values that are perceived as the user's domain knowledge. The user can save the action items as a transformation profile to apply in the future resulting data from the workflow.

The interactive, visual exploration panel provides a view that recommends visualization methods to users based on the data type and format. Users may also independently select relevant visualizations from the palette of available visualization methods. This palette is also extensible to allow users to add highly tailored visualizations for specialized data or analysis tasks. This latter feature is provided

in recognition of anticipated unconventional visualization requirements for different varying use cases **(R7)**. To support interactive, visual exploration, we modularized the exploration panel based on the use case. As such, the visual exploration panel for each use case inherits the common visualizations and includes (optional) custom visualizations. For example, to support the sensemaking in one use case (discussed in Section 3.5.2), we implemented the interactive custom scatter plot shown in Figure 3.10. The inherited visualization library includes line charts, standard scatter plots, parallel coordinates, box plots, heat maps, geospatial maps, and tabular data presentations.

3.4.3 Distributed Analysis System: VIFI

To evaluate VAF, we integrate VAF with two DAS: a simple file-based DAS and the Virtual Information Fabric Infrastructure (VIFI) DAS. In this section, we describe the latter DAS which serves as the foundation of most of our VAF evaluation activities.

VIFI [123, 127, 128, 132] is a DAS that enables analyses across distributed, fragmented data without the movement of massive data. Within VIFI, analyses migrate to the distributed data and only derived data â e.g., result sets â migrate from the data hosts. VIFI supports research and analysis in multiple domains including astronomy [132], earth science [123], and sustainable human-building ecosystems (SHBE) [128]. The current implementation of VIFI consists of the following components: Portable Analytic Containers, Registry Services, Orchestrator, User Node, and Data Sites. Each is described briefly in the following.

Portable Analytic Containers (PACs): A PAC is a lightweight virtual machine, called a container, that hosts software, libraries, and the operating system needed by end users to analyze data. A PAC can receive and execute analysis programs (e.g., scripts) if the required programs are not already contained in the PAC. Leveraging container technology (e.g., Docker [157, 158, 159, 160]). A PAC is portable to migrate and execute on heterogeneous host platforms. A PAC facilitates reusability

by hosting and utilizing different analytical libraries and programs pulled from shared repositories (e.g., Docker hub [169]). Container technology enables the movement of analytics rather than the movement of data; thus, alleviating problems related to the transfer of big data. PACs offer a number of affordances for distributed analytics: i) they can be easily transmitted over the network due to their limited size; and ii) they simplify analytics development for inexperienced users. The VIFI infrastructure is scalable as it enables the integration of various VIFI nodes at different sites. The ability for VIFI workflows to access fixed sites allows VIFI to cooperate with non-open-source resources, assuming that a VIFI user has the proper credentials. Currently, VIFI researchers are extending VIFI to use Singularity [170, 171, 172] to run on High Performance Computing (HPC) clusters at different sites.

Registry Services: Distinctive PACs are stored, searched, utilized and shared through Registry Services. Currently, VIFI uses Docker hub [169] to implement the Registry Services. We expect future VIFI versions to incorporate additional services to advance the download and transfer times of PACs.

Orchestrator: The Orchestrator automatically coordinates workflow (i.e., task) execution across multiple VIFI sites (i.e., distributed datasets). Each analysis step in a workflow is implemented by a script running in a PAC at a data site. Although initial VIFI implementations used NiFi [163, 161] as its orchestrator, current implementation use RESTful APIs to improve orchestrator customizability.

User Node: The user node is the means by which users interact with the VIFI framework. The user node provides a UI, communication, and basic computation capacities.

Data Site: Data Sites are locations in the VIFI infrastructure where distributed, fragmented data reside. Each VIFI Data Site interacts with the Orchestrator (i.e., NIFI and/or RESTful APIs) and runs PACs (e.g., by Docker Swarm [173]). VIFI uses Docker Swarm to execute parallel analytics. Each Data Site runs a VIFI server

supported by a configuration file that configures hosted data sets and log files at this site. **Metadata Server:** The Metadata Server stores and lists gathered metadata about datasets at each Data Site. These metadata are used to support data discovery (foraging).

Crawler: The Crawler is used by the Metadata Server to automatically collect metadata at each Data Site.

Watchdog: The Watchdog updates the Metadata Server when modifications to the metadata are detected at any of the associate Data Sites.

VIFI workflows are either launched from the command line interface of the VIFI server running at each Data Site or via the User Node. The VAF reported in this paper functions as the VIFI User Node for the use case evaluations reported in the following section that used VIFI.

VIFI workflows can be launched from the command line interface by `python vifi.py --sets [list of VIFI sets that should be initiated on current VIFI Node] --vifi_conf [local VIFI Server configuration file]` on each VIFI Node that participates to the workflow. The `--sets` option specifies which VIFI Sets should be launched on the current VIFI Node, while the `--vifi_conf` option specifies the VIFI Server configuration file. The VIFI Server configuration file contains information about all VIFI Sets hosted at the current VIFI Node including the data (e.g., XML, NETCDF, CSV, etc.) path for each set, the data exposed name to be discovered and used by end-users, the allowed container images to run within each VIFI Set and their locations, and other configurations. Additionally, the VIFI Server configuration file contains information related to logging, variables specifications used for each VIFI Set directory structure and reserved file names. The VIFI Server configuration file is a YAML file that can be extended for future development of VIFI. Each VIFI Node will be ready to accept users' requests using RestAPI. The workflow designer can use different VIFI Sets hosted on remotely distributed VIFI Nodes. Thus, the end-user

will be able to analyze accessible data at each VIFI Node using the designed workflow without moving the data to the end-user location.

3.5 Use cases

To evaluate the affordances of our VAF, we implemented the framework leveraging the VIFI DAS. As part of our evaluation, we present two use cases: one from the earth sciences and the other from the SHBE domain [174]. Guided by researchers from these domains, we implemented workflows that integrated the researcher's analytic scripts. The earth sciences use case included two workflows and the SHBE use case included three workflows.

Implementing a new use case in VAF includes three steps. First, we use the workflow/script management view to create the new use case in the use case management middleware repository. This step generates a unique use case key and associates it with a user-specified name. All subsequent workflows and their execution results will be associated with this key. The user also specifies the DAS data site(s) or hosts that will be leveraged by the workflows.

The second step involves reviewing the DAS configuration data. For VIFI, these data, stored in the `conf.yml` file and submitted to VIFI during task execution, specify the constraints that govern VIFI communications.

The third and final step for use case creation involves verifying that proper infrastructure constraints are satisfied. For example, proper firewall and security standards need verification with the organizations that will be hosting the VIFI infrastructure. Once a use case is created, workflows may be specified and executed, and results may be explored. In the following sections, we illustrate VAF through workflows from each evaluation use case. Figure 5.3 denotes the technologies we leveraged for our implementation.

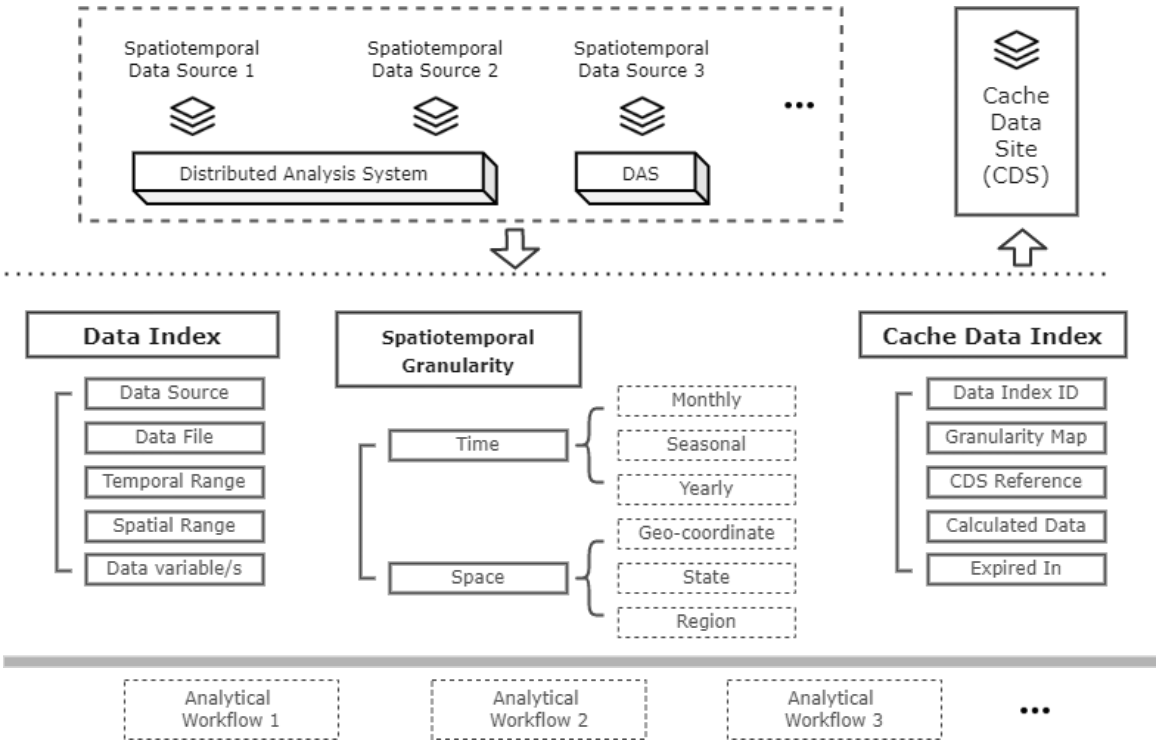


Figure 3.7: The data mining architecture for distributed spatiotemporal data. The deciding factors for resulting data storage are based on usage frequency, analysis runtime, and required storage memory.

3.5.1 Earth Science: Exploring Climate Projections

We used our VAF, leveraging the VIFI DAS, on NASA Earth Exchange published downscaled climate projections (NEX-DCP30) [175]. The United States National Climate Assessment (NCA) [176] reports the future projections of the various climate variables from NEX-DCP30 to assess changing climate scenarios [21, 13]. Recognizing its importance, the NASA Earth Exchange project released NEX-DCP30 data (observed and projected) that contain monthly averaged precipitation and temperature data for the contiguous US from 1985 to 2099. The projection data are stored in Network Common Data Form (NetCDF) [52] format and provide access to the projection output for 36 climate models [175].

To perform demonstration evaluations of our VAF integrated with VIFI, we worked with a NASA climate scientist to develop workflows for analyzing NEX-DCP30. These

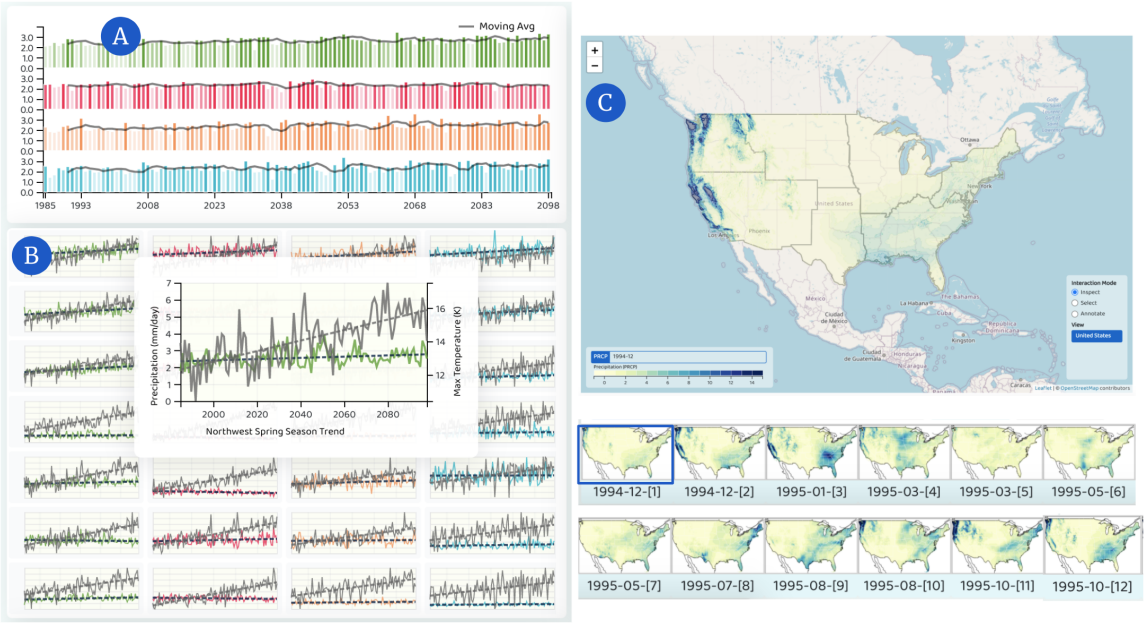


Figure 3.8: The visual analytic interface for the earth science use case, leveraging VIFI. Interactive geospatial visualization and trends for seasonal regional temperature and precipitation assist climate scientists in their analytic tasks.

workflows extracted the NetCDF data files and summarized monthly averaged spatiotemporal data for interactive, visual exploration as illustrated in Figure 3.7. The first workflow executes data extraction analyses based on user-provided parameters, such as projection model(s), climate variable(s), and year(s). An analytic script uses these parameters to find the corresponding NetCDF data and extracts geospatial contours for each month of the given year. The script and workflow configuration were authored and stored in the middleware using the Code Editor. The configuration file identifies the dataset (e.g., NEX-DCP30) and links via the middleware to authorization credential required for execution. In fact, the workflow configuration file contains all of the required parameters to execute this workflow. Hence, each time users execute a workflow, they update the parameters in the configuration file to extract the projection model of interest. The resulting data are formatted as GeoJSONs [177], subsequently stored in the middleware repository (e.g., an S3 bucket). Once data extraction is complete, the user can visualize and interactively explore the

results as shown in Figure 3.8C. Recall that the VAF visualization library provides a generic map view that renders the geospatial contour visualizations. The geospatial navigator in Figure 3.8C is coordinated with the geospatial view, rendered using a configurable slider built in the visualization library.

The second earth sciences workflow summarizes the spatiotemporal climate projections from NEX-DCP30 for exploration and analysis. This workflow contains multiple analytic scripts to summarize data from different perspectives while using different statistical techniques. Multiple scripts are included in this workflow since they share similar analysis goals. Users can reconfigure the workflow to use different scripts based on preference and interest. Workflow results contain monthly, seasonal, and yearly summaries of precipitation and temperature grouped by season and region. We created custom visualizations for this workflow as depicted in Figure 3.8. The requirement for this custom visualization was identified and co-designed by the participating climate scientist. Figure 3.8A shows multiple bar charts, sharing similar axes, illustrating the mean precipitation from 1985 to 2098, for each season. Figure 3.8B provides small multiples of precipitation and temperature trends for the 21st century. Each small multiple denotes a region and season correspondingly from top to bottom and left to right. In this use case, the custom visualization can be used for exploration and analyses independent of the script that configures the workflow.

3.5.2 SHBE: Light Switching in Smart Buildings

The SHBE domain is a multidisciplinary field that explores the interplay of human behaviors and the built environment with the goal of a more sustainable future. Multiple workflows have been explored in collaboration with SHBE researchers. For space consideration, we highlight just one of these workflows to illustrate how more complex workflow designs are supported and enabled by VAF. The analytical purpose of the highlighted SHBE workflow is to explore the use and efficacy of Artificial Neural Networks (ANN) for the prediction of light on-off switching probabilities for the work

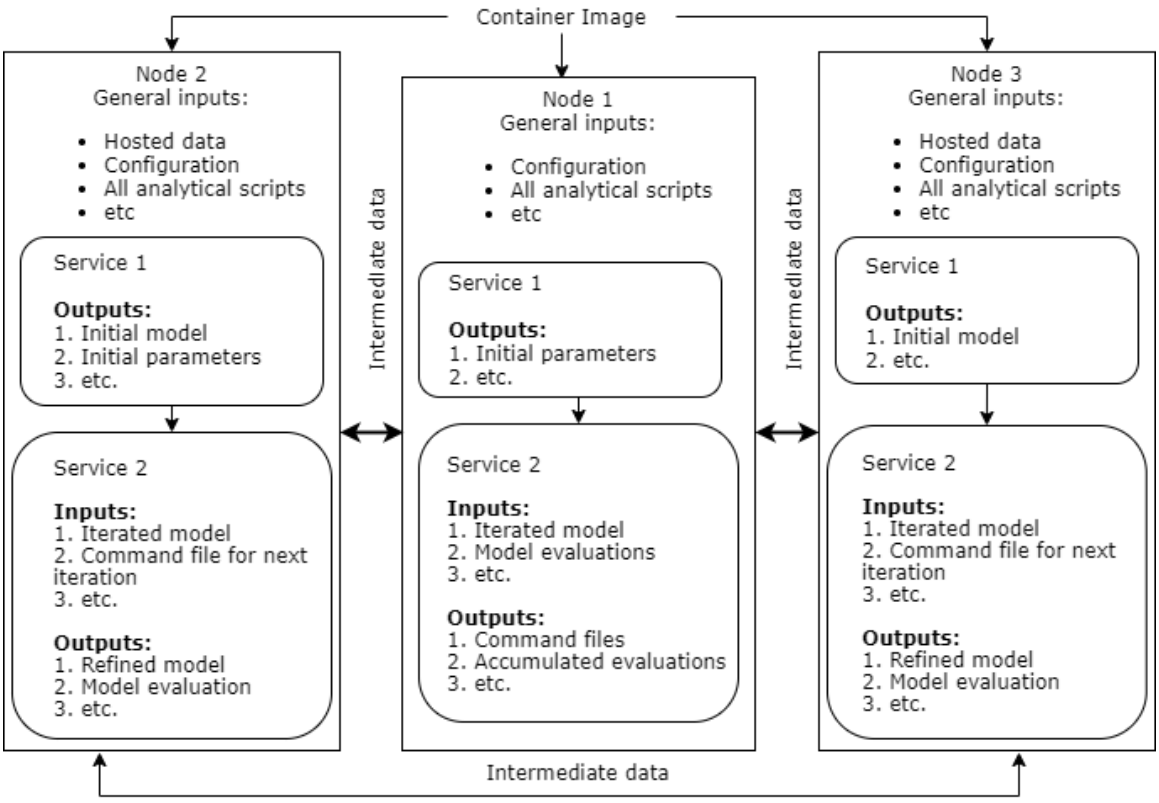


Figure 3.9: Workflow implementation for SHBE light switch on-off probability in smart buildings.

area illuminance in a smart building as shown in the interactive VAF visualization presented in Figure 3.10. To illustrate the complexity of the analyses, we summarize the workflow implementation in VIFI below.

As shown in Figure 3.9, the workflow involves analysis over three distributed datasets at three different VIFI Data Sites. The data at each VIFI Data Site is used by the ANN model for training and prediction. The third VIFI Data Site collects the updated ANN model and determines whether further model refinement is required using any of the other 2 VIFI Nodes. Thus, the third VIFI Node sends a different command file to each Node to specify what to do in the next step (e.g., fit the ANN model using existing data, use the ANN model to make predictions, etc.). Finally, when the third VIFI Node decides that the model is "good enough", the stopping condition is reached. The VIFI Orchestrator terminates the workflow and

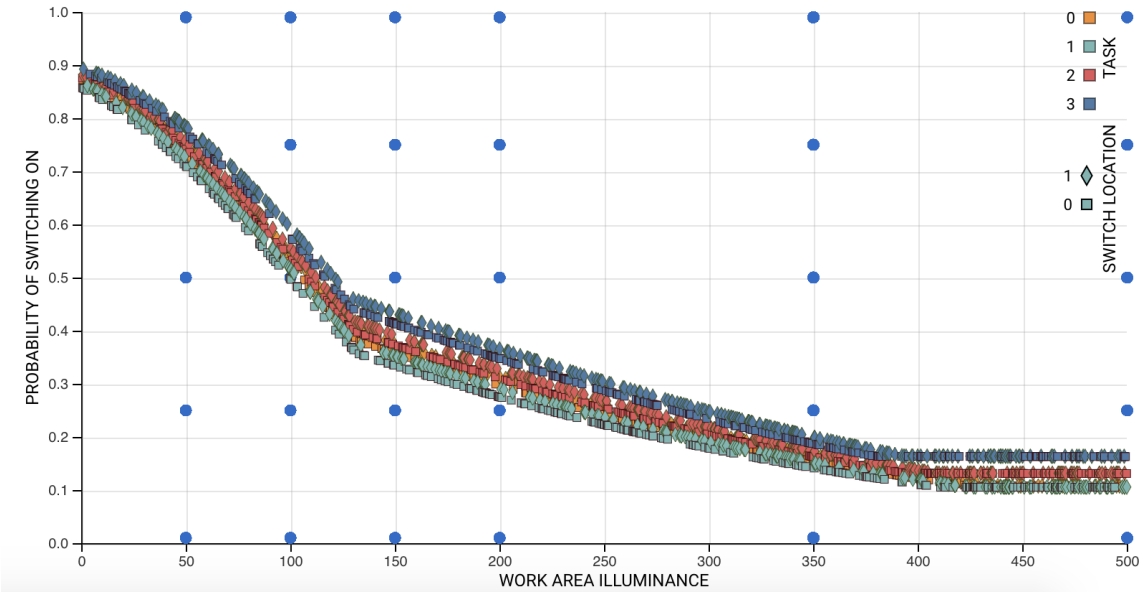


Figure 3.10: The SHBE use case result exploration for the smart building workflow. The scatterplot illustrates the light switching on-off probabilities based on the work area illuminance using the ANN model.

results are returned to the VAF middleware.

The ANN model, as well as other intermediate results, are sent between the VIFI Data Sites using RESTful API-based VIFI Orchestrator. The RESTful API is also used by each VIFI Data Site to accept incoming requests for analyses from users launching workflow. Similar to the earth science use case, each request for analyses contains the required scripts, parameters, and workflow configuration. The configuration contains important information for proper workflow execution including the dataset(s), PAC(s), and input parameters as well as operation settings such as where to send intermediate and final results, whether to keep a local copy of the (intermediate) results for further analysis, whether to add timestamps to results for potential time-series analyses and other similar settings. In this workflow, analysis at each VIFI Data Site consists of two steps (or scripts). The output of each step is stored locally and transferred to other VIFI Data Sites for further processing. The first step in each Data Site in this workflow executes only once but its output is used in multiple subsequent steps at this and other Data Sites. In other words, the initial

ANN model is created at one of the VIFI Data sites as step one and it is used to predict outcomes and/or to train models at subsequent steps. Thus, it may execute any number of times until it is decided that the ANN model is âgood enoughâ and the workflow is terminated. As mentioned previously, VAF supports the specification of the workflow and renders the output as a scatter plot as depicted in Figure 3.10. This interactive visualization is customized so that square and diamond-shaped glyphs denote switch-on and switch-off operations while color is used to denote independent workflow runs. The visualization describes the probability of light switch behavior for work area illuminance.

3.6 Discussion

We presented a VAF for DAS to assist the data owners, researchers, and analysts to manage the infrastructure and conduct analysis through a web-based graphical UI. We have reviewed several distributed analysis systems such as XSEDE [124], SciServer [109], and VIFI [123] to identify the design requirements to resolve the requirement for the user to directly access the server, manage the access control from the application layer, and facilitate the user to explore the result using interactive visualizations.

We identified 7 implementation requirements that satisfy the design requirements to develop a web-based graphical UI for DAS. An interface for preparing the analytic scripts, configuring the workflow, and running the workflow in DAS sites resolves the requirement for the users to directly access the DAS servers. The middleware orchestrates the transactions between the UI and DAS. Moreover, the middleware manages the authentication and authorization from the application layer to reduce the workload of data owners. The workflows executed by the users through the UI are queued in the middleware database. The middleware communicates with the DAS sites to decide when to push the queued tasks and provides runtime and error logs to the UI that help the user to monitor the progress of the task. Finally, the visual exploration panel produces interactive visualizations to explore the resulting

data from the analytic scripts.

We demonstrated the UI that satisfies the design requirements and illustrates the implementation requirements of our proposed VAF. The UI consists of 3 main panels - workflow management, task management, and visual exploration. The workflow management provides access to a hierarchical file structure (Figure 3.4A), a component for creating or updating analytic scripts (Figure 3.4B), and a terminal (Figure 3.4C) to provide raw streaming logs. The middleware serves RESTful APIs to synchronize the UI with the DAS site on user interactions. The task management panel provides status updates for the running workflows, and overall runtime progress, and allows the user to either re-run the workflow or explore the result (Figure 3.5). The visual exploration panel familiarizes the user with the data (Figure 3.6), and perceives their preferences to produce a set of interactive visualizations.

We implemented VAF in two use cases from the earth science and SHBE domain. We leverage VIFI [123] DAS to implement 2 workflows from earth science and 3 workflows from SHBE. These workflows were initially configured and executed through a command line interface. The users were required to access multiple servers including the data sites to run their analyses. In contrast, after the initial configuration and setup of VAF, the users are not required to access the distributed servers to create, update and run their analytical scripts. The pre-configured visual exploration panels for respected workflows assisted the analyst users to explore the result without any effort in creating interactive visualizations.

Nevertheless, we identified a few limitations of VAF based on our implementation experience. Our framework complies only with the DAS that provides RESTful APIs. We plan to address this issue by developing a generic RESTful API and deploy at the DAS sites to comply with more distributed systems. The workflow configuration from the UI requires a learning curve for the users to be familiar with the configuration keywords We plan to provide a better interface with more readable

labels and input validations for the configuration items which would ease the user with workflow configuration. We understand our visualization library lacks the use case-specific visualization and interaction requirement to explore the results, which required workflow designers' effort to preset the visualizations. We plan to create more input scopes for the users to inject their domain knowledge to influence the visualization recommendation [178].

CHAPTER 4: INTERACTIVE GEOSPATIAL VISUALIZATION

Exploratory user interactions in contour-based visualizations create approaches to visual analysis that are noticeably different from the perspective of human cognition. As such, visualization researchers have introduced diverse interactive approaches to improve system usability and enhance human cognition in the exploration of complex and large-scale spatial datasets. However, further research is needed to better understand and quantify the potentials and benefits offered by user interactions in contour-based geospatial visualizations designed to support exploratory analytical tasks. In this paper, we present a contour-based interactive geospatial visualization designed for exploratory analytical tasks. We also report on the design and development of a web-based user interface that facilitates quantitative user intervention studies of interactive contour visualizations. We conduct a crowd-sourced user intervention study (N=62) that analyzes the impact of interactive features in the contour-based geospatial visualizations in support of exploratory analytical tasks. Our results evidently show that the interactive features lead participants to perform significantly better compared to static visualizations. Moreover, the quantitative analysis of participants' performance and observations from their interaction data provide a deeper understanding of utilities in the interactive features from the perspective of spatial data extent, map layout, task complexity, and user expertise. Finally, we comprehensively discuss our findings that serve as guidelines for future design and implementation of interactive features for contour-based geospatial view in support of case-specific analytical tasks.

4.1 Introduction

Geospatial visualizations are increasingly being used to support the analysis of spatial variables in many research domains including atmospheric science [12, 13, 14], meteorology, [90, 39, 179] and urban planning [15]. In addition, scientists and analysts in those domains leverage contour maps in intricate geographical analysis such as un-

derstanding landform shapes, mountain elevations, ocean depths, climate projections, etc. In recent years, visualization researchers and domain experts are collaborating with greater frequency, employing cross-domain research efforts to identify innovative visualization and interaction techniques that help domain experts extract unknown events and hidden patterns in the geospatial data [63, 44]. However, our literature review includes such collaborations in the earth sciences domain [119, 27, 180, 85] and revealed that many researchers rely on traditional static depictions of geospatial data in contour visualization to perform common visual analysis tasks [181, 116].

For instance, international assessment reports are a critical resource for understanding climate change, assessing societal impacts, and supporting effective decision-making. The National Climate Assessment (NCA) reports published by Intergovernmental Panel on Climate Change (IPCC) are projections from climate models [182]. The assessment reports provide both direct foretelling of physical indicators of potential future climates and indirect information on societal impacts where contour maps play an essential role in visually interpreting downscaled climate projections. As the spatial resolution of climate models becomes finer, model performance is generally improving, providing better representations of extreme weather events [116], the hydrological cycle [183], and influential land surface processes [184]. However, traditional approaches often leverage static contour depiction of high-resolution spatial data [113], leaving the useful underlying events and insights from finer spatial resolution. In this context, we hypothesize that user interactions with the contour-based geospatial visualizations enhance the efficacy of such analyses; but, it is clear that measuring these benefits for exploratory analyses remains a difficult task.

User interactions and visual representations often organize, structure, and segment spatial data in ways that enable user exploration from abstract overviews to focused inspections. Exploratory user interactions in geospatial visualizations create approaches to visual analysis that are noticeably different from the perspective

of human cognition. In recognition of these differences and the opportunities that they afford, visualization researchers have introduced diverse interactive approaches to improve system usability and enhance human cognition in the exploration of complex and large-scale datasets [94]. The importance and value of user interactions in supporting exploratory geospatial analysis are well-accepted phenomena. However, quantitative evaluations of the interactive features in such systems are lacking, despite the qualitative evidence that demonstrates their value. In particular, further research is needed to better understand and quantify the potentials and benefits offered by user interactions in geospatial visualizations designed to support exploratory analytical tasks.

To quantitatively evaluate the interactive features in contour-based geospatial visualization, the comparative artifacts are interactive and traditional static maps. User interactions in visualization applications are often categorized based on their support for different analytical tasks [27]. Hence, to conduct a quantitative evaluation, it is important to define a set of exploratory tasks and comparative visual artifacts. Previous research categorizes the exploratory geospatial analysis from different perspectives, including event identification, comparison, ordering, and logical reasoning [185, 186, 187]. Moreover, selecting proper datasets, ones that obtain adequate information to resemble these analytical tasks is consequential for the evaluation outcome.

In this paper, we report our findings from a crowd-sourced user intervention study that quantitatively evaluates interactive features in geovisual environments supporting exploratory analytical tasks. We leverage knowledge from our literature review [188, 186] and collaboration with the climate and environmental scientists [44, 113, 114] to select seven exploratory visual analyses to investigate the usability of the interactive features in contour-based geospatial visualizations. To conduct the study, we utilize geospatial data containing temperature records for a tri-state area [189] and precipitation data for the contiguous U.S. [21]. Data collected from the user study is

statistically analyzed to understand quantitatively the impact of the user interactions in performing the identified exploratory analytical tasks. Our results show that interactive geospatial visualizations improve correctness in task performance compared to static geospatial visualizations. In particular, the interactive features significantly improve task performance in the identification of spatiotemporal differences from the geospatial contours. This chapter presents following contributions:

1. We present a contour-based interactive geospatial visualization that supports the users performing exploratory analytical tasks. Geospatial visualization consists of numerous features that allow users to inspect, select, annotate, and filter the geospatial data.

2. We designed a web-based user interface that facilitates a user intervention study under an exploratory contour map visualization environment. We conducted a crowd-sourced user intervention study to scrutinize the interactive features in the contour-based geospatial visualization.

3. We discuss participants' interaction behavior and report diagnostic analysis of their performance that serves as guidelines to design and evaluate case-specific interactive geovisualization.

4.2 Related Works

In this section, we summarize the previous research in visualizing the geospatial data that enables users to conduct exploratory visual analyses. In addition, we review the methods and techniques in evaluating the usability of the geovisual environments.

4.2.1 Geospatial Visualization

Geospatial data are traditionally categorized into three types: raster data, vector data, and network data [9, 190]. Raster data gather a large set of pixels where each pixel corresponds to a specific geographic location. These data often illustrate discrete

information such as terrain type, as well as, continuous information such as temperature or land elevation. In contrast, vector data contains information such as the location of a movie theater, the path of roads, rivers, and the area of a country, river, or national parks [9]. Vector data is often represented by points, lines, or polygons. Moreover, network or graph data illustrates the spatial connectivity among the nodes in the vector data such as road transportation and navigation [191]. In addition to spatial features, dimension-impacting factors like time are frequently associated with the geospatial data [192, 193].

Visualization researchers have introduced many geovisual environments for various data types with analysis requirements, and use cases. The power of geospatial data through geographic information systems for mapping and analytics has been realized beyond earth science-related domains. Such domains (e.g., traffic systems, energy systems, health informatics, entertainment) heavily rely on the location or geographical data [194]. Choropleth and geospatial maps are popularly leveraged to visualize the data attribute containing spatial dimensions [88, 38, 3].

The static geospatial depiction is frequently utilized where users visually estimate information from the traditional map views in form of raster images. NASA built Panoply, a software package, to extract high-dimensional climate datasets and visualize them as geospatial contours [97]. QGIS [195] is another widely used open-source cross-platform software for visualizing geospatial data in a diverse format (e.g., GeoJSON, NetCDF, Shapefile) as a mesh layer. In addition, toolkits such as MATLAB [103], NCAR [105], and GrADS [104] are still commonly used to visualize the geospatial data.

Moreover, we reviewed transformations of geospatial visualizations to accommodate the multi-dimensional attributes. CDAT provides a software infrastructure and platform for analyzing multidimensional data [196, 99], popularly used for distributed analyses of geospatial data. In addition, CDAT offers visualization and control sys-

tems that interact with Python to visualize the geospatial data [196]. GIS4WRF [197] is another toolkit for processing and visualizing weather data and forecasting NetCDF in an interactive single graphical environment. MeteoInfo is a desktop application that provides a geospatial visualization platform for multidimensional meteorological data [198].

4.2.2 Evaluating Geovisual Environments

Tools and applications employed for geospatial visual analysis are often developed for use case-specific needs. Therefore, qualitatively evaluating the usability of those tools is mostly sufficient with the domain experts' or corresponding users' feedback [66, 113]. In our review, we only found a handful of literature that quantitatively evaluated geovisual environments with users from a diverse range of expertise and the effect on their performance. Koua et al. [186] conducted a user study with geographers, cartographers, geologists, and environmental scientists as participants where they evaluated the usability of different visualization methods for geovisual analyses. Their study included visualization methods such as parallel coordinates, and self-organizing heatmap representation against the geospatial map to evaluate users' performance in 10 exploratory analytical tasks. Zahan et al. [199] conducted a user study to compare four design techniques for contour line stylization to explore the multivariate geospatial data. Nagel et al. [200] introduced a tabletop map visualization that enables users to explore geospatial networks interactively. Moreover, Mahmood et al. [201] presented a mixed reality geospatial visualization method that fosters remote collaboration and knowledge sharing for sensemaking tasks. We understand the exploratory geospatial task classification presented by Koua et al. [186] is relevant to our work as we design our user study for quantitative evaluation.

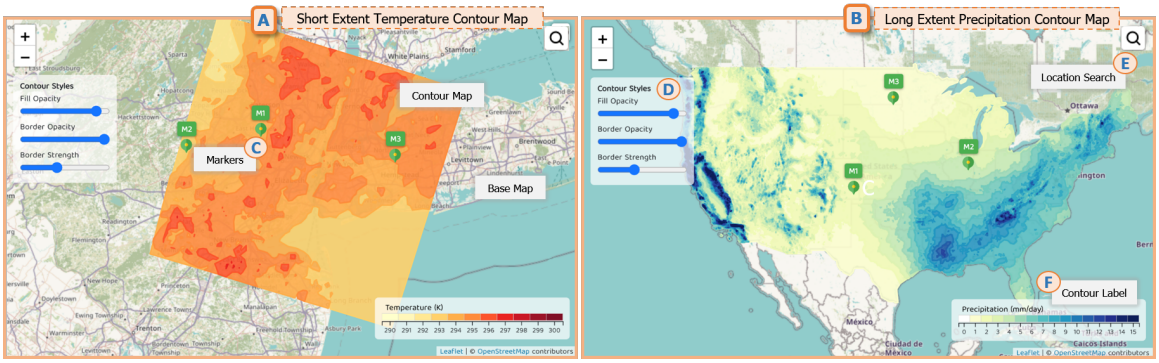


Figure 4.1: Overview of our implementation of an interactive geospatial visualization presenting the contour map of (A) temperature intensity in New York City and (B) precipitation intensity in the contiguous U.S. recorded at a certain time. The interactive features in the map are intended to support exploratory visual analysis providing underlying information about the contour regions.

4.3 Contour-based Interactive Geospatial Visualization

For our user study, we developed an interactive contour-based geospatial visualization that includes multiple coordinated visual and interactive components. These components are 1) a geospatial map, 2) a contour area distribution view, and 3) interactive features including, but not limited to, inspection, selection, exploration, toggle layers, draw markers, control visibility, and location search. The visual components alone can illustrate the geospatial data (often in the GeoJSON format) and basic mapping to the geographic location without leveraging the interactive features. We utilized Leaflet.js to visualize the geospatial contour transformed into GeoJSON.

4.3.1 Geospatial Map

The geospatial map presents a multi-layer view that includes a base map, a contour map, and spatial masks. In geographic information system (GIS), the base map refers to a collection of geographic features that illustrates the background scene for a location. For instance, we use OpenStreetMap [202] as a base map to show geographical features (e.g., boundaries, rivers, and highways). The contour map provides the topographic view of a location, where is the area sharing the equal value clustered

with lines or polygons [203]. In Fig. 4.1, for example, the contour maps present A) temperature intensity for New York City, and B) precipitation intensity for the contiguous U.S. Additionally, the spatial masks draw boundaries for the contiguous US regions, that provide a visual segment of the area specific (regional) intensity. While this multi-layer version of the geospatial contour map obtains scope to accommodate interactive features over a traditional single-layer bitmap, we still refer to this map as a static map in the user study.

4.3.2 User Interactions

We implemented several interactive features over the traditional geospatial map to transform the contour into an interactive map. For the interactive version, we keep the map layout unique to the static map, only including components that enable user interactions, to ensure the comparison between static and interactive maps is focused on the interactive features instead of the visualization design. The user interactions are designed to facilitate the human cognition process and categorized based on their capabilities in supporting the exploratory analysis within geospatial data [94]. The interactive features are designed to perform the exploratory analysis task discussed later in Section 4.4.2. To implement the interactive features, we have gathered knowledge from our previous research [113, 114, 44] and literature review [36, 40, 12]. We included a set of interactive features such as inspection, selection, annotation, filter, encode, and view synchronization.

The boundaries within the contour create multiple polygons which we identify as the contour regions. In the interactive map, the interface can calculate the total number of regions belonging to unique intensity. To calculate the contour area distribution, we group the contour map based on the area that shares an equal value/range of intensity. Then, we aggregate the area of all the contour regions grouped by unique value/range of intensities to identify the contour area distribution.

Inspect: This interaction allows the user to move the mouse cursor over a contour

region, then, the map highlights all the associated contour regions on the map. It also pops up a tooltip that provides additional information about the interacted contour region. The information on the tooltip includes the intensity of the interacted region, the number of contour regions occupied by the intensity, and the contour area distribution by those regions.

Select: This interaction allows the user to enable or disable all the contour regions of an intensity. The user can click on the contour labels (Fig. 4.1F) to select the events of interest. The contour area and regions are re-calculated based on the updated selection.

Annotate: We implemented the annotation feature where the user can place markers on the map to explore detailed information or compare it with other locations or regions. A right mouse click on the point of interest puts a marker on the map (Fig. 4.1C). The view allows multiple markers in different locations. A left click on the mouse shows additional information as an inspection of the marked location and illustrates the comparison with the other marked places.

Filter: This interaction essentially provides search capabilities within the contour. Search assists the user to locate an area/bound on the map that might be difficult to find manually. To search an area, the user may click on the search icon (Fig. 4.1E) placed at the top right on the contour map, type the area name and choose the one from the options. A marker will be placed on the searched and selected location. The searched markers function just as the other annotation markers placed on the map.

Encode: The map view encoding supports the user controlling the visibility of the contour on the geographical boundary. Contour styles allow controlling the transparency or strength of the contour area and borders to be able to better co-relate an event with its geographical location.

Synchronize: This interaction caters to the comparison between two contours, providing exploratory features to identify the potential differences for points of inter-

est. The interactive features discussed above are synchronized in this setup, meaning the user interaction with a contour map has a complementary effect on the comparative map. In this study, we present these comparative maps as 'Snapshot Left' and 'Snapshot Right'.

The interactive also provides navigation interaction such as zooming and dragging the viewport. However, we have not included those interactions in our analysis as we intend to focus only on the usability of the user interactions with the contour regions. An implementation of the contour-based interactive geospatial map shown in Fig. 4.1, demonstrates the interactive features selected for this user study. The participating features are decided based on their potential support to the exploratory analysis tasks discussed later in Section 4.4.2. That entails, there are plenty of scopes remaining to include more interactive features that have usage scenarios for specific cases.

4.4 Study Design

While contour maps are extensively used to visualize the spatial changes of the data variables, scientists and researchers often leverage additional analysis toolkits to understand the depth of information. With technological and infrastructure advancement, geospatial data, nowadays, contains high-dimensional and high-resolution variables which are critical to understanding geographical characteristics. Hence, we established research questions and hypotheses to measure the effect of interactive features in contour maps. We select a subset of exploratory analysis task types that are previously leveraged in evaluating exploratory geovisual environments [185, 186]. We designed analysis tasks for this user study, derived from selected task types that are common among climate and environmental scientists [186, 113, 199, 114].

4.4.1 Research Questions & Hypotheses

We recognize the potential of interactive contours on a geospatial view in visually analyzing high-resolution geospatial data. Therefore, understanding the utility of interactive features in geospatial visualization is significant to outline its potential to the full extent. To measure the impact of the interactive features in the contour-based geospatial maps, we breakdown our core research question into the following questions:

RQ1. Can interactive contour maps assist users in visually performing analytical tasks with better correctness in their estimation? Additionally, how does it affect performance in terms of task completion time?

Design & Hypothesis: Participants will be prompted with multiple sets of questionnaires from different geospatial contour maps. The contour maps are divided into two groups where one group is presented with the interactive features and others with the static illustration.

H1: The users' performance in exploratory geospatial analysis tasks can be benefited from the interactive features in the map. The ability to explore characteristics of contour regions with user interactions can lead them toward a better estimation of the information. Moreover, we hypothesize that the participants require more time to complete tasks since there are more elements to interact with and cognitive processing.

RQ2. Do the interactive features on the map hold a significant difference against static maps in terms of users' performance in various types of exploratory analysis tasks?

Design & Hypothesis: We leveraged a taxonomy of exploratory geospatial analysis to design the tasks (further discussed in Section 4.4.2) and sketched the potential workflows in context to visual analysis. The participants will be prompted with multiple sets of unique analysis tasks from the classification. The tasks in a set

are designed around a contour map and provide the full extent of interactive features to trial with a group of participants. Another group of participants performed similar tasks with the static version of the same contour maps.

H2: We hypothesize that the interactive geospatial maps improve users' performance in the complex exploratory analysis tasks in comparison to the static maps. The users are expected to perform better in tasks where perceiving the relationships among multiple data points or contour regions. We suspect that would require users to spend more time completing the tasks in exchange for better correctness. However, we hypothesize that the overall individual performance in terms of correctness per time would be improved significantly.

H3: We also hypothesize that interactive maps narrow down the participants' performance gap between exploring the single geospatial data and comparing multiple data instances. Comparative exploration between multiple maps is expected to be more challenging than exploring a single geospatial map. As opposed to static map, interactive maps synchronize the user the interactions, therefore, obtains the potential to extract more correctness in multiple map exploration tasks.

H4: Interactive maps elevate the performance more for novice users who are not familiar with geospatial analysis compared to experienced users. Performance measure in this context includes both correctness score and task completion time. We still expect experienced users to achieve a correctness score significantly better than novice users. However, the interactive maps could lead novice users to an improved correctness score per time, meaning the time requires to interact with the map would have a greater impact on novice users than the experienced users with geospatial analysis.

We intend to measure the likeliness of utilizing each individual interactive feature in different exploratory analysis tasks. We report the usage of the interactive features and analyze the underlying patterns in the usage. The patterns are quantified based

on the time differential of performing the tasks and their usage sequence.

4.4.2 Task Specification

We design this study to conduct a data-driven investigation to evaluate whether interactive features in the geospatial visualization benefit performing exploratory geospatial analysis tasks. To design the user study, we relied on the literature review to find the exploratory analysis tasks for the geo-visualization environment. Exploratory geospatial operations and analysis tasks often fall under the identification of clusters, the relationship between data points, comparison, and analysis of relevance to the geographical location. Numerous pieces of literature have translated these tasks to visualization operations [204, 185, 188]. However, these visualization operations are supported by multiple visual artifacts including geospatial view, parallel coordinates, matrix representations, line plots, etc. [186, 187].

In this study, we adopted the comprehensive list provided by Keller et al. [185] and later leveraged by Koua et al. [186] to conduct a usability study for exploratory geovisualization. We designed our analytical tasks wrapped around these operations, which we identify as the following 7 task types: *1) Associate, 2) Categorize, 3) Cluster, 4) Distinguish, 5) Identify, 6) Locate, and 7) Rank.* ***Associate*** tasks involve users in finding similarities between data points based on the characteristics of attributes, geographic location, and patterns of events. ***Categorize*** tasks require defining all regions on the display by the boundaries. The regions are usually categorized by common features or spatial positioning. ***Cluster***-type tasks require users to identify different groups of similar data points or regions within the geospatial data. In these tasks, the user identifies sparse data points or regions that share common attributes to form unique groups within the data. ***Distinguish*** analytical tasks require the users to recognize the difference or variations between the spatial events. These differences often include geospatial area or location and values of the same attribute. ***Identify*** tasks demand establishing the relationship between data attributes based on their

shared characteristics. ***Locate*** requires recognizing certain data points or events on the map. In this task, we aim to evaluate whether the map supports the user in locating the facts or a range of values. ***Rank***-type tasks require feature ordering based on the data attributes, clusters, or geographic locations.

Based on the number of data points involved in the exploration, we classified the exploratory task types into simple and complex tasks. Simple tasks include *distinguish*, *identify*, and *locate* types of spatial exploration where the users engage with not more than two data points. In these tasks, the user is either aware of the event and querying the location of where it happened or vice-versa. In contrast, complex tasks include *associate*, *categorize*, *cluster*, and *rank* type of spatial analysis where the user explore the relationships among multiple data points. The users generally identify the data similarities and mapping based on a point of interest regardless of location or event.

We designed the tasks for each type from the point of view of a single instance of geospatial data exploration and multiple geospatial data comparisons. The synchronize interaction implemented in the map is expected to facilitate the users in comparing the spatial events between multi-maps. Such cases are frequently studied in identifying the temporal changes in spatial data. While the single map obtains the potential to support the exploration of an instance of geospatial data, we intend to measure the support that interactive features provide in analyzing the spatiotemporal differences. Therefore, we designed a set of tasks that account for both single and multi-map analysis and exploratory task types as outlined in Table 4.1. We understand that the analytical tasks in Table 4.1 can be related to more than one task type, however, we assigned the type based on the most prominent resemblance of the analysis.

Table 4.1: The participants performed these exploratory geovisual analyses from 7 unique task types. These exploratory tasks are designed for both single and multiple map layouts to evaluate the usability of the interactive features in a geovisual environment.

Task Type	**Single Map**	**Multiple Maps**
Associate	Based on the area covered which two *[variable/marker]* intensities below are the most similar?	Find the *[variable/marker]* intensities from both snapshot that cover the *[largest/smallest]* area.
Categorize	What is the *[area/intensity]* (difference) for the *[highest intensity/largest area]* recorded on *[contour/Location X]*?	What is the area difference of *[variable]* *[intensity]* between Snapshot Left and Right?
Cluster	How many *[unique/total]* contour regions are visible on the map?	How has the number of *[unique/total]* contour regions changed from 'Snapshot Left' to 'Snapshot Right'?
Distinguish	What is the *[area/intensity]* difference between the highest and lowest *[variable]*?	How does the *[highest/lowest]* intensity shift from 'Snapshot Left' to 'Snapshot Right'?
Identify	What is the *[intensity/difference between]* *[variable]* in Location *[X]* (and Location *[Y]*)?	What is the intensity difference between Location *[X]* in 'Snapshot Left' and Location *[Y]* in 'Snapshot Right'?
Locate	What is the *[highest/lowest]* intensity on the *[contour/marked area]*?	What is the *[highest/lowest]* intensity *[between both snapshots/among all markers]*?
Rank	Order the *[markers/intensities below]* from *[high to low/low to high]* *[intensity/area]*.	Order the snapshots based on the *[larger/smaller]* area covered by the *[lowest/highest]* *[variable]* intensity.

4.5 User Study

4.5.1 Geospatial Datasets

In this study, we scrutinize the contour-based geospatial visualization under two scenarios, presenting the long and short extent of the geospatial data. Hence, we observe whether the geographical extent and data granularity has an impact on the quantitative evaluation of the interactive features. We present geospatial visualization utilizing the data generated from two different sources:

1. Hourly-averaged temperature forecast for the tri-state area based on the Weather and Research Forecasting (WRF) model published by Coastal Urban Environmental Research Group (CUREG) [189], as the sample illustrated in Fig. 4.1A.

2. Monthly-averaged precipitation projections for the contiguous U.S. (CONUS) published by NASA Earth eXchange (NEX) where downscaled climate projections (NEX-DCP30) are generated in different emission scenarios such as RCP 2.8, RCP 3.5 and RCP 4.5 [21], as the sample illustrated in Fig. 4.1B.

Temperature data (D_T) for New York City cover a short spatial extent, thus, allow to experiment with interactive features such as inspection, search, marking, visibility control, and synchronized interaction. In contrast, precipitation data (D_P) for the Contiguous U.S. cover a relatively large extent, therefore, allow us to experiment with additional interactive features such as region/state selection, data filtering, and viewport shifting. To complement the repetitive measures outlined in Table 4.2, we arbitrarily included six temperature and precipitation data instances in this study, recorded at distinct times.

4.5.2 Study Setup

We designed a custom web-based interactive interface to conduct the user intervention study [1]. The user study interface is developed on React.js using D3.js

[1]Link to the user study interface: https://geospatial-study.vercel.app/

Table 4.2: Distribution of the 8 sets of tasks that accounts for the repetitive measures such as the spatial extent, visualization methods, and map layout. Each task set includes a task from the classification of the exploratory visual analysis tasks.

Map Data Context	**Map Data (Task Set)**	**Participant Group 1**	**Participant Group 2**
Short Spatial Extent (NY Temperature)	Single Map - D_T1	Static Maps	Interactive Maps
	Multi Map - [D_T2, D_T3]		
Long Spatial Extent (CONUS Precipitation)	Single Map - D_P1		
	Multi Map - [D_P2, D_P3]		
Short Spatial Extent (NY Temperature)	Single Map - D_T4	Interactive Maps	Static Maps
	Multi Map - [D_T5, D_T6]		
Long Spatial Extent (CONUS Precipitation)	Single Map - D_P4		
	Multi Map - [D_P5, D_P6]		

and Leaflet.js for interactive visualizations, Django for the API server, and MongoDB for storing the application and user data. Additionally, we reviewed literature that conducted crowd-sourced studies to evaluate the visualization components [205, 200, 206, 207]. Our review provided us with a well-grounded idea about participant recruitment and exclusion criteria, training requirements, and study procedures.

First, the interface provides an overview to the participants about the study procedure, assignment details, and interaction data to be captured and asks for their consent. Then, the interface collects basic demographic information from the participants such as level of education, gender, expertise with geospatial data analysis, familiarity with data visualization, and daily computer usage. Next, the interface forwards the participants to a training window to describe the interactive features and familiarize them with the features in visual analysis. The training window includes 12 sections where we explain the contour map, geospatial data, and interactive features. In addition, the training window prompts a couple of sample questions after

providing detailed explanations for each interactive item to verify the participant's understanding. There is no pass or fail in the training phase. If participants select an incorrect answer, the interface interactively indicates the verdict and guides them to the correct answer. This interactive guidance and hints are predominantly implemented for the training and later disabled in the user study part. Once the training is complete, the interface will redirect the participant to the user study window.

In the user study, the interface prompts 8 sets of tasks (divided into two sections) where each set is associated with a map setup presenting different geospatial data. Each task set includes one question from each of the 7 task types which sum to 56 questions (8×7) for the study. Table 4.2 provides a fair distribution of the map and task types in counterbalancing the samples and taking repetitive measures. While the participants perform the tasks by interacting or observing the map visualization, the interface captures their observations, task completion time in seconds, and the use of interactive features associated with questions. After completing the study, the participants get a completion code from the interface which they submit to the Amazon Mechanical Turk (MTurk) [208] portal to earn their incentive.

4.5.3 Participants & Recruitment

We conducted the study virtually online. The participants are recruited from MTurk. We posted a human intelligent task (HIT) in the Mturk portal that links to our user study interface. To ensure the cleanliness of the collected data, we included participants who hold more than 95% approval ratings on the MTurk portal and are identified as master workers. As per our pre-registration, we excluded the participants who failed to correctly answer the engagement questions and completed the study in less than 8 minutes. These requirements left us with 62 participants (25 females and 37 males) for this study. We performed an a-priori power analysis using G* power application [**?**]. The result suggests that a sample size of 62 estimated achieves power over 0.994. We asked participants about their academic qualifications, their

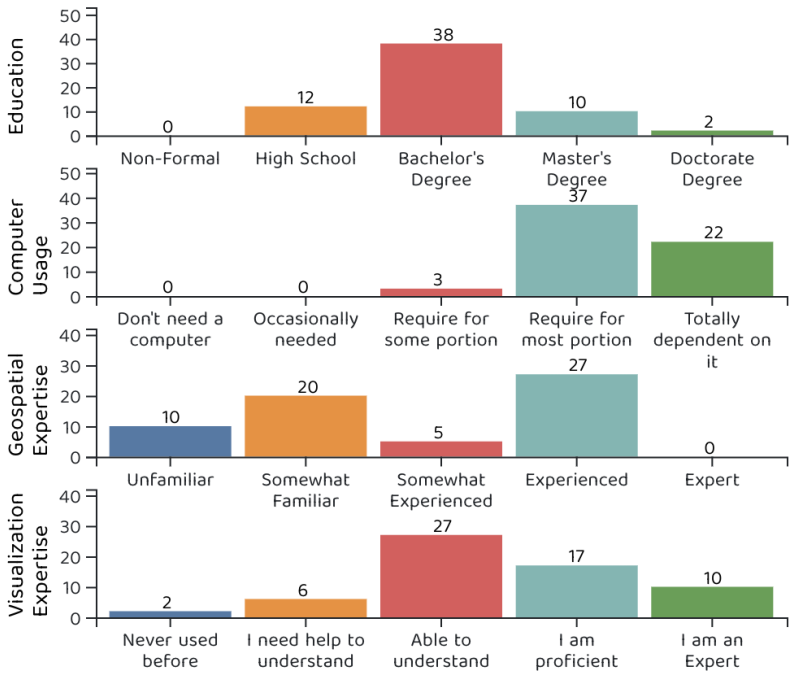


Figure 4.2: Summary of 62 participants' educational qualifications, computer usage, geospatial expertise, and visualization expertise.

daily computer usage, and how they rate their expertise in perceiving visualization and analyzing geospatial data. For the computer usage, visualization expertise, and geospatial analysis expertise, the participants identified themselves from the groups shown in Figure 4.2. We found that 80% participants have at least a bachelor's degree and all of them use computers for their daily tasks. They have identified themselves at various levels of expertise with visualization and geospatial analysis as Figure 4.2 suggests. Participants earned a $4 incentive for their effort and contribution to our user study. They took an average of 8.6 minutes (*SD=7.99*) to complete the training and an average of 29.14 minutes (*SD=19.75*) to complete the study.

4.5.4 Evaluation Metrics

Evaluations of geospatial visualizations span numerous qualitative and quantitative dimensions [119]. In this study, we focused on quantitative metrics to evaluate the interactive features of the geospatial map. In particular, to assess the dimen-

sions of the exploratory analysis tasks and validate our hypotheses, we consider the following quantifiable factors - participant's task completion time, the correctness of their performed task, and the use of visualization or interaction techniques to perform the task. We utilize these data, to measure a quantitative analysis with respect to static and interactive maps as well as how the task type factors in the result. The performance metrics are designed to report a quantitative analysis that addresses the research questions stated in the earlier section.

Dependent & Independent Variables: In this study, we considered two dependent variables: 1) the task completion time, and 2) the correctness of the performed tasks. For the independent variables, we considered the interactive features on the map (visualization method) as discussed in Section 4.3.2, and the exploratory task types as discussed in Section 4.4.2.

4.5.5 Analysis Methods

We leverage two-way and one-way repeated measures analysis of variance (ANOVA) [209] at a 5% significance level as our primary method to test the effect of the independent variables on different exploratory task types. In our analysis, we apply two-way ANOVA to evaluate the performance metrics based on two independent variables that include participants' correctness score and task completion time. Additionally, to present the estimates we calculate the lower and upper bounds at a 95% confidence interval (CI). To present the difference between the visualization methods, we report the p-value and the error bound mean (EBM) for the task types. Further, we plot the participants' performance in an interactive geospatial map against the static map in a normal distribution. Both the participants' correctness and task completion times are taken into consideration to interpret the performance. In addition, we quantitatively rank the usability of the interactive features in the visual analytics interface for the participating task types. We expect that it would be beneficial for visualization researchers in the future to employ different interactive features based

Table 4.3: Two-way ANOVA results for correctness and time required with visualization method as static vs. interactive maps and exploratory task types.

-	**Factor**	***F***	p	η_p^2
Correct	Vis × Task	(6, 366) = 2.908	**.006**	.046
	Vis	(1, 61) = 114.186	< **.001**	.652
	Task	(6, 366) = 107.459	< **.001**	.638
Time	Vis × Task	(6, 366) = .892	.513	.015
	Vis	(1, 61) = 5.755	**.020**	.088
	Task	(6, 366) = 7.698	<**.001**	.114

on the intended analysis tasks.

4.6 Results

In this section, we apply statistical methods to testify our hypotheses and validate the research questions raised in Section 4.4.1. To test the hypotheses, we have considered two measures from data collected in the user study: the correctness score of the performed tasks and task completion time. Correctness is measured based on the participant's ability to estimate what the tasks require. The completion times are recorded by the UI based on the time elapsed before the participants completed each task.

4.6.1 Effect of User Interactions

The two-way ANOVA result that is presented in Table 4.3 shows in terms of overall correctness, there is a significant interaction effect ($p = .006$) between the static ($M = .537$) and interactive ($M = .647$) visualization methods and the exploratory task types. The simple effect of the visualization method ($p < .001$) and the task types ($p < .001$) are also significant. For the task completion time, there is no interaction effect ($p = .448$) between the visualization methods and the task types. The main effect of the visualization method ($p = .060$) and the task types ($p = .129$) are also

not significant.

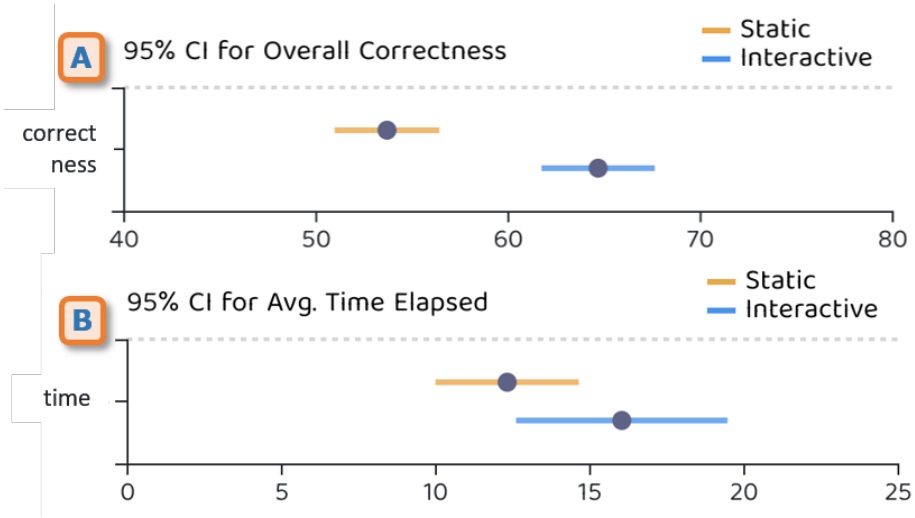


Figure 4.3: The error bound mean (EBM) for overall correctness score and task completion time estimate at 95% CI. The result shows participants' (n=*62*) correctness significantly improved in interactive maps.

Figure 4.3 presents the 95% CI of the overall correctness and time elapsed to perform the analysis tasks using both the static and interactive geospatial maps. Interactive geospatial maps ($EBM = [.62, .68]$) evidently assist ($z = 5.47$) in better correctness than static maps ($EBM = [.51, .56]$). However, the participants spent a relatively greater time (z=1.79) with the interactive maps (EBM=[12.63, 19.49]) than the static maps (EBM=[10.01, 14.66]) while performing the analysis tasks. This result verifies **H1** that the interactive maps assist the users in producing a significantly accurate result, however, requires an insignificantly greater time than the static maps in leading to the observation. This makes sense as the interactive features allow to explore the underlying information about the map while the participants interacted with the contour regions. We also quantified how the interactive features improved the individuals' performance, analyzing by normally distributing their performance fluctuation. The result suggests that the participants are expected to perform consistently (R^2=.903) better using the interactive features in the geospatial map.

Table 4.4: Correctness and completion time results of the tasks from exploratory task types using the static and interactive maps.

	Correctness			**Completion Time**		
Task Type	**F (1, 61)**	**p**	$\eta_p{}^2$	**F (1, 61)**	**p**	$\eta_p{}^2$
Associate	24.257	**<.001**	.285	5.988	**.017**	.091
Categorize	13.897	**<.001**	.186	8.601	**.005**	.125
Cluster	20.902	**<.001**	.255	2.436	.124	.039
Distinguish	2.531	.117	.042	2.750	.102	.044
Identify	4.923	**.030**	.075	8.263	**.006**	.121
Locate	3.279	.058	.058	.839	.363	.014
Rank	9.445	**.003**	.134	.018	.893	.000

4.6.2 Interaction Effect on Exploratory Tasks

To test **H2**, we measured the statistical significance of the interactive features in the performance. We applied ANOVA and compute the mean differences (MD) in pairwise correctness and error bound mean (EBM) for correctness score and time estimates between static and interactive maps. The results of ANOVA comparing the visualization methods in terms of task completion time and correctness are presented in Table 4.4.

Associate: In associate-type tasks, participants' correctness shows a significant difference ($p < .001$) between the static and interactive maps. The pairwise correctness difference ($MD = [.371, .908]$) leans onto the interactive map. The correctness score suggests that participants performed substantially better using the interactive features ($M = 1.855$, $EBM = [1.621, 2.089]$) in the map compared to static maps ($M = 1.177$, $EBM = [.990, 1.364]$). That being said, the difference in time elapsed between the visualization methods to perform the associate tasks are also significant ($p = .017$). The participants spent significantly more time interacting with the interactive maps ($M = 125.237$, $EBM = [91.438, 159.036]$) compared to the static maps ($M = 91.673$, $EBM = [68.676, 114.671]$).

Categorize: The difference between static and interactive maps is significant

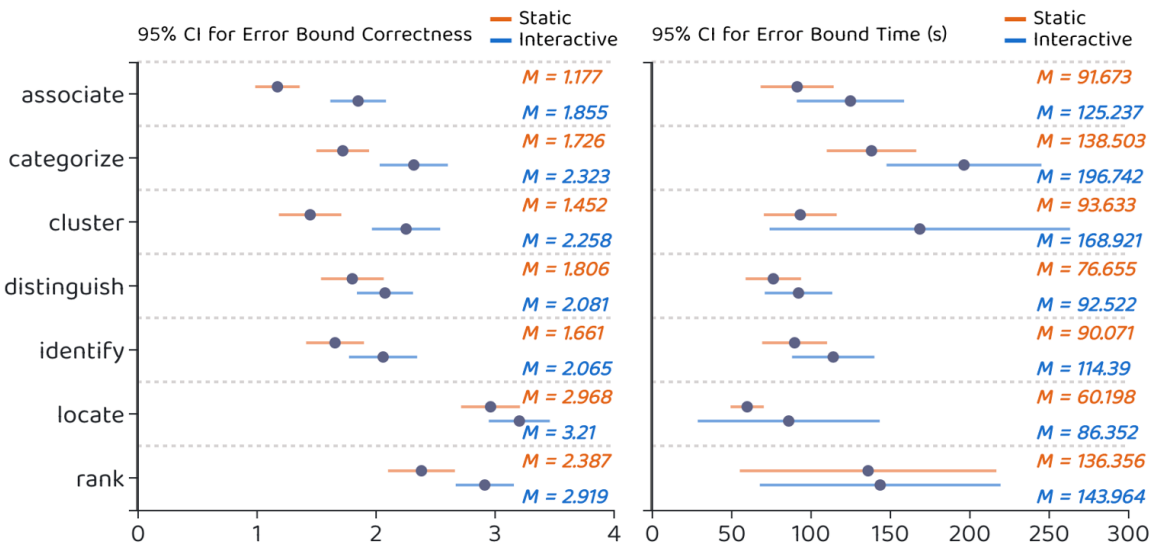


Figure 4.4: The participants' error bound mean (EBM) correctness score and task completion times. While the completion time is often faster with the static maps, the correctness scores are consistently better with the interactive maps.

($p < .001$) on the basis of participants' performance matrices such as time spent and correctness. The correctness score in the interactive maps ($M = 2.323$, $EBM = [2.036, 2.609]$) and the static map ($M = 1.726$, $EBM = [1.504, 1.947]$) shows that the interactive features provide the participants better support in performing the categorize-type tasks. The pairwise difference observed in the data also indicates better results ($MD = [.265, .915]$) in the interactive maps. While the correctness improved in the interactive maps, the participants also tend to spend significantly more time ($p = .005$) performing the tasks in interactive maps ($M = 196.742$, $EBM = [147.961, 245.524]$) compared to the static maps ($M = 138.503$, $EBM = [110.308, 166.698]$) as shown in Figure 4.4.

Cluster: We observed statistically significant differences ($p < .001$) between static and interactive maps in participants' correctness while performing the clustering tasks. The pairwise correctness difference suggests the participants achieve better correctness using the interactive maps ($MD = [.445, 1.162]$) which are also supported by the correctness scores estimates for the interactive maps ($M = 2.258$, $EBM =$

$[1.971, 2.545]$) compared to the static maps ($M = 1.452$, $EBM = [1.189, 1.714]$). The difference between the participating visualization methods in terms of time spent is not statistically significant ($p = .124$), however, it still suggests that the participants spent more time in the interactive maps ($M = 168.921$, $EBM = [74.280, 263.562]$) than the static maps ($M = 93.633$, $EBM = [70.681, 116.585]$).

Distinguish: The correctness difference between the visualization methods is statistically insignificant ($p = .117$) for the distinguish tasks. The pairwise correctness difference also indicates the insignificance ($MD = [-.031, .654]$). The correctness score estimates still suggest that the participants performed better with the interactive maps ($M = 2.081$, $EBM = [1.845, 2.316]$) than with the static maps ($M = 1.806$, $EBM = [1.543, 2.070]$). Similarly, the difference in time spent between these visualization methods is also insignificant ($p = .102$). The static maps ($M = 76.655$, $EBM = [59.187, 94.124]$) lead to perform slightly faster than the interactive maps ($M = 92.522$, $EBM = [71.263, 113.781]$).

Identify: The participants' correctness illustrates a significant difference ($p = .030$) between the visualization methods for the identify-type tasks. The pairwise correctness difference ($MD = [.011, .743]$) is inclined to favor the interactive geospatial map. The correctness scores indicate that the participants achieved better results with the interactive maps ($M = 2.065$, $EBM = [1.778, 2.351]$) compared to the static maps ($M = 1.661$, $EBM = [1.418, 1.904]$). The difference in time spent between these visualization methods is also statistically significant ($p = .006$). The participants spent more time with the interactive maps ($M = 114.390$, $EBM = [88.458, 140.323]$) than the static maps ($M = 90.071$, $EBM = [69.620, 110.521]$) while performing the identify-type tasks.

Locate: Interestingly, there is no significant difference ($p = .058$) between the visualization methods in terms of participants' correctness. The pairwise correctness difference ($MD = [-.024, .515]$) compared to the static map also indicates there is no

statistical significance. The correctness score estimates suggest that the performance is marginally better with the interactive maps ($M = 3.210$, $EBM = [2.953, 3.466]$) than with the static maps ($M = 2.968$, $EBM = [2.720, 3.215]$). The difference in time spent between the visualization methods is also statistically insignificant ($p = .363$). The task completion time estimates do not follow the trend of other task types where participants completed tasks faster in static maps. It is evident that the participants required similar times to complete the locate-types tasks in static maps ($M = 60.198$, $EBM = [49.739, 70.657]$) whereas interactive maps suggest a wide elapsed time range ($M = 86.352$, $EBM = [29.037, 143.667]$).

Rank: We observed a statistically significant difference ($p = .003$) in participants' correctness between the visualization methods. The pairwise correctness shows the interactive maps extract better correctness ($MD = [.350, .515]$) from the participants. The correctness score at 95% CI suggests the participants' performance is consistently better with the interactive maps ($M = 2.919$, $EBM = [2.675, 3.164]$) than the static maps ($M = 2.387$, $EBM = [2.106, 2.668]$) for the ranking tasks. The time differences between the visualization methods are not significant ($p = .893$). The CI suggests that the participants spent an estimated equal time completing the ranking tasks in both the interactive ($M = 143.964$, $EBM = [68.157, 219.771]$) and static ($M = 136.356$, $EBM = [55.567, 217.144]$) geospatial maps.

4.6.3 Interaction Effect on Multi-maps Exploration

The ANOVA result shows evidence that the interactive maps narrow down the participants' correctness score gap to a slender margin between single and multi-map exploration (**H3**). For associate tasks in static maps, the correctness difference between single ($M = .790$, $EBM = [.20, 1.563]$) and multi-maps ($M = .387$, $EBM = [.091, 1.099]$) is statistically significant ($p = .001$) whereas, in interactive maps, the difference becomes insignificant ($p = 0.715$) between single ($M = .952$, $EBM = [.327, 1.819]$) and multi-maps ($M = .903$, $EBM = [.256, 1.72]$). Cluster (static $p =$

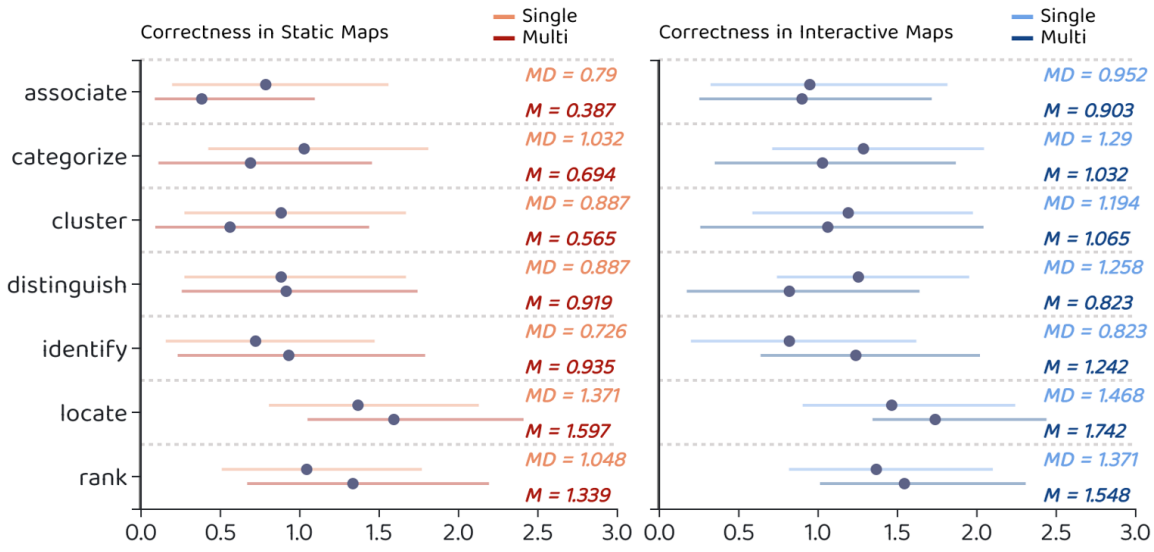


Figure 4.5: Correctness comparison between single and multi-maps using static (left) and interactive (right) geospatial visualization.

.017; interactive $p = .368$) and rank-type (static $p = .023$; interactive $p = .125$) tasks show similar observations to a different degree. While doing cluster tasks in static maps, the correctness scores in single maps ($M = 1.032$, $EBM = [.353, 1.872]$) are significantly better than the multi-maps ($M = .694$, $EBM = [.114, 1.459]$) whereas, in interactive maps the difference between single ($M = 1.194$, $EBM = [.590, 1.979]$) and multi-maps ($M = 1.065$, $EBM = [.262, 2.046]$) are insignificant.

The participants explored the single map instance with better correctness scores than the multi-map where identify ($MD = -.21$), locate ($MD = -.23$), and rank-type ($MD = -.29$) tasks are proven to be odd. For example, in the case of rank-type tasks, the correctness difference between single ($M = 1.371$, $EBM = [.821, 2.105]$) and multi-maps ($M = 1.548$, $EBM = [1.016, 2.312]$) becomes insignificant in interactive maps, however, for both static and interactive maps, the participants performed better in the multi-maps compared to the single maps. The correctness score mean along with the error bound means (EBM) are presented in Figure 4.5. Regardless of the performance gap between single and multi-maps, the participants' correctness score improved in all task types in interactive maps, except for the distinguish-type

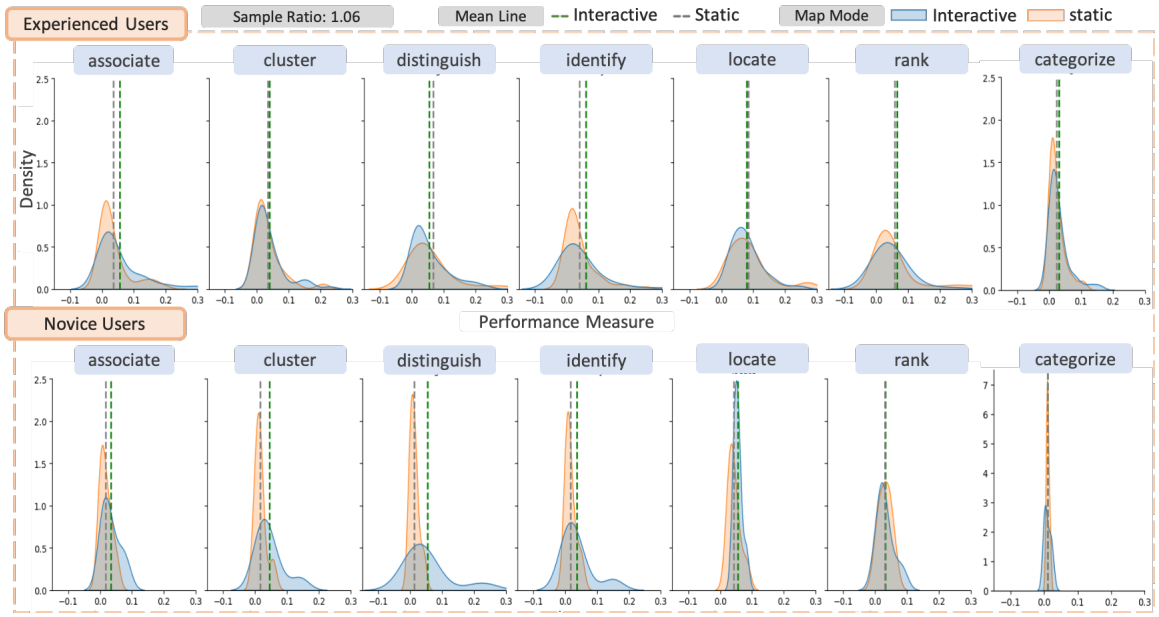


Figure 4.6: Comparison between novice and experienced participants' performance suggests that novice users excel with interactive features significantly more than experienced users.

tasks. These findings reaffirm both the hypotheses **H2** and **H3**.

4.6.4 Performance Elevation by Geospatial Expertise

To test hypothesis **H4**, we classified the sample space of participants into two groups - novice and experienced, based on their provided expertise in geospatial analysis. The participants who identified themselves as *'unfamiliar'* or *'somewhat familiar'* with geospatial analysis are classified as novice users. In contrast, participants who identified themselves as *'somewhat experienced'*, *'experienced'*, or *'expert'* are classified as experienced users. As stated in **H4**, we take both correctness score and completion time into account to verify if the novice users gained more performance boost using interactive contour-map than the experienced users. The sample ratio between these two groups is 1.06 which is fairly equal, as Fig. 4.2 suggests.

To measure the performance while taking both correctness and task completion time into account, we calculated the participants' correctness per unit of time in the 7 exploratory analysis task types. Fig. 4.6 illustrates task-wise density plots to

separately compare the performance measure of both novice and experienced users in interactive and static maps respectively. We draw dashed lines in Fig. 4.6 to represent the performance mean for static (gray) and interactive (green) maps. The experienced users' overall performance is significantly better than the novice users. In addition, we previously verified with **H1** and **H2** that all participants tend to spend significantly more time in interactive maps to complete the tasks. Here, the mean performance difference between interactive and static maps suggests that interactive features elevate the novice users' performance more than the experienced users, thus verifying **H4**. While spatial data categorization tasks are found to be particularly difficult for novice users, the performance seems to excel in other complex task types such as associate, cluster, and rank.

4.7 Discussion

The experiments in this user study outline the benefits of the contour-based interactive geospatial map in various exploratory analysis tasks. Exploratory data analysis (EDA), as John W. Tukey stated, is an act of *looking at data to see what it seems to say* [210]. Therefore, EDA is essential for building the hypothesis by initial investigation of identifying patterns or discovering anomalous events in the data. With that considerations, we implemented a set of interactive features for the geospatial map that facilitates the users to interactively explore underlying information about the contour regions on the map. The interactive features presented in this paper are a concept built around the presumed requirements of the participating exploratory analysis tasks, drawn from previous research [113, 114, 44] and an abundant number of literature review [36, 40, 12]. Nonetheless, the user study design and experiments illustrate a systematic quantitative approach to measure the usability of interactive features for an exploratory geovisual environment.

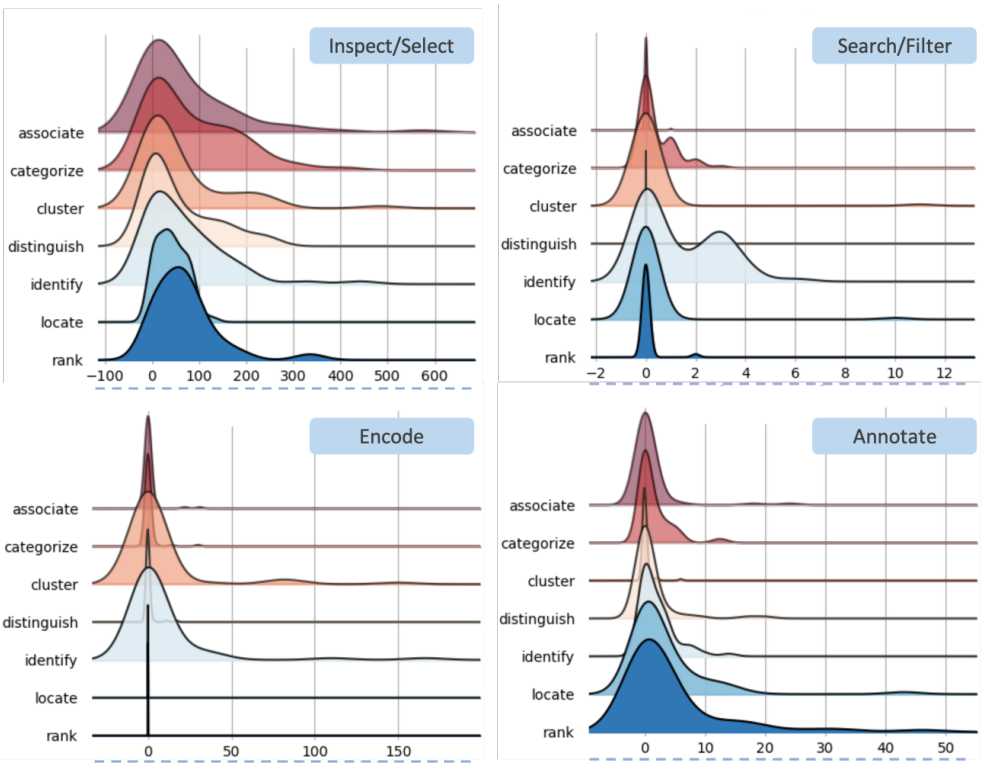


Figure 4.7: The four most frequently used interactive features in the participating ($t = 56$) exploratory tasks. The plotted range of usage frequency is categorized by the task types and bootstrapped with a 95% confidence interval.

4.7.1 Observations from the User Interaction Logs

We analyzed interaction logs from the perspectives of usage frequency, usage pattern over the task completion time, and usage sequence of each interactive feature. We present these metrics based on the exploratory task types. To understand participants' interaction usage, pattern, and sequence, we classified the interaction based on the number of mouse/keyboard operations (step) required to perform the feature.

Identifying the frequently used interactive features. Analyzing participants' interaction data with the interactive contour map, we observe that inspecting/selecting the contour regions is the most popular followed by annotating, searching/filtering, and controlling the contour styles. This finding is reasonable as inspect/select is the single-step interaction that put up the basic information about the contour area, as opposed to other multi-step interactions such as search, annotation,

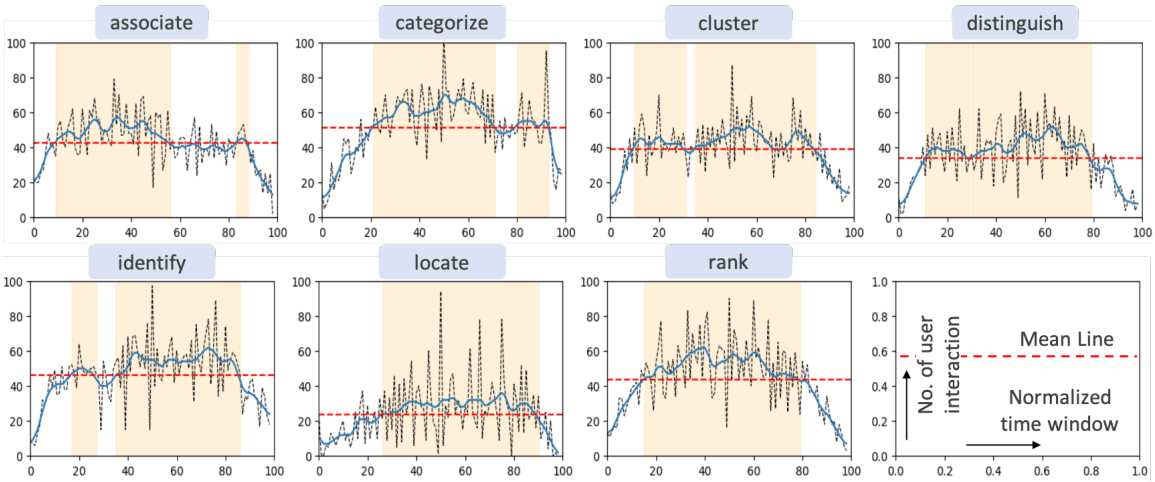


Figure 4.8: Task-wise user interaction pattern based on the normalized task completion time window. The highlighted area represents the time segments where participants' interactivity exceeds the mean.

and encode. While summarizing the data, we identified a substantial number of unintended interactions in the data. For example, the participants are only required to navigate the cursor to inspect the contour regions, therefore, to reach the area laying deep inside the map border, the user interacts with other unintended contour regions as well. Before the analysis, we trimmed off these unintended interactions by implementing a time threshold. Hence, we determine the regions where participants spent reasonable time perceiving the information provided on the tooltip or widgets. Fig. 4.7 shows the credible usage summary of the four most frequent interactive features while performing ($t = 56$) exploratory analysis tasks (as distributed in Table 4.2) by each participant.

Reflecting cognitive process with interaction patterns. We also observed the participants' interactive behavior during the task completion time window. Therefore, we normalized the window size and plotted the user interaction (Fig. 4.8) to find any foreseeable pattern. Inactivity in terms of user interactions during the initial and final segments in the window indicates participants' cognitive processing to understand the task and reach an analytic decision. We draw a red line for each task type in Fig. 4.8 to indicate the mean interaction usage across the normalized time

window. The highlighted area shows the window where the participants' interactivity is above the mean. It also provides the opportunity to estimate the segments where the participants are seemingly cognitively active. We understand participants are more interactive in complex tasks such as associate, categorize, and rank, as well as, spending comparatively more time cognitively processing the data points relations. Moreover, We observed participants relatively less interactive in the distinguish and locate type tasks compared to other types. This finding is also reflected in the correctness score, as the difference is insignificant between static and interactive maps in these task types.

Perceiving the reasoning of interaction usage. We observed from the interaction data of the complex tasks that participants inspected/selected various distinct contour regions a great many times before reaching a decision. Since the associate, categorize, and cluster tasks require establishing the relationships among a set of distinct intensities, therefore, they reasonably interacted back and forth to reach an analytical decision. Annotations are mostly used in the locate, rank, and associate-type tasks to mark/inspect spatial data points. These interactions include participants' self-made annotations and interaction with predefined markers on the maps. Annotations are particularly useful in the ranking tasks as the marker widget provides comparison insights with other annotations, hence, supports in deciding the order of the contour regions. We also observed that annotation interactions are generally performed followed by inspect/select and filter interactions. Filter interactions are frequently utilized while performing the identify and categorize-type tasks. These interactions include searching based on geolocation. Identify tasks required to extract information based on geographical characteristics. Therefore, the participants realized the search feature to eventually annotate or inspect intensities based on the locations. Comparative information provided by the annotated locations is continuously updated as the user interacts with other contour regions. Consequently, proven

effective for categorization and clustering as these tasks require identifying the regions with similar variability. Encoding features are mostly used in cluster and identify-type tasks. The contour encodings are understandably useful in the cluster tasks, as it often cases required to identify the unique regions on the contour map. The identify tasks involve the geographical characteristics that usually reside in the base map, hence, the participants' controlled styles correlate with the contour map in these tasks. Therefore, changing the weight and opacity of the contour lines and region's area can be beneficial in visually performing these tasks.

4.7.2 Interpretation of Results

We established two research questions to evaluate the usability of interactive contour maps in the participating seven types of exploratory analysis tasks. The results suggest that our initial hypothesis (**RQ1**) is correct as the participants achieved improved correctness using the interactive contour maps as opposed to the traditional static contours as shown in Fig. 4.3. Next, we hypothesize that the interactive maps should reflect a significantly improved correctness score in the participating exploratory analysis task types. The participants are also expected to spend significantly more time with the interactive contour maps compared to the static maps. The ANOVA result (Table 4.4) shows that participants' correctness scores in the associate, categorize, cluster, identify, and rank improved significantly (**RQ2**) while we also observe statistically insignificant improvement in the distinguish and locate-type tasks. Additionally, the result shows that the participants consumed more time to perform the tasks with the interactive maps where the differences in associate, categorize, and identify-type tasks are significant as shown in Fig. 4.4.

As the task distribution presented in Table 4.2, we designed the repeated measures based on the spatial extent (*short vs. long*), map layout (*single vs. multi map*), geospatial expertise (*novice vs. experienced*), task types (*7 EDA*), and map interactivity (*static vs. interactive*). Thus, we analyzed the fluctuations in participants'

correctness scores accounting for these individual factors in interactive maps against static maps. We understand that the participants' correctness score difference in the static maps is significant between the single and multiple maps, especially for the associate, categorize, cluster, and rank-type tasks (Fig. 4.5). In contrast, the score difference in the interactive maps between the single and multiple maps is statistically insignificant except for the locate and distinguish-type tasks. While we have not found any statistically significant interaction effect on the correctness score and the exploratory task types based on short versus long extent maps, our analysis shows evidence that the interactive mode brings the user performance close. We also measured user performance based on their reported geospatial expertise, considering both correctness score and completion time. While the experienced users consistently performed better than the novice users in both modes, the performance gaps are certainly closer between them in the interactive mode (Fig. 4.6), especially in complex tasks such as associate, cluster, and rank.

Quantitative evaluation in this study shares similarities with the previous studies [186, 40, 200] in terms of task design, participant recruitment, and analysis methods. However, the results from our experiments provide a unique understanding of the usability of the interactive features from different perspectives, as opposed to previous geovisual user studies, where only different visualization components [186], display designs [201], and visual stylization [199] are evaluated for visual analysis tasks. Hence, this study provided scope to not only quantitatively measure participants' analytical performance but also their interaction behavior to report both descriptive and diagnostic analysis from the usage data. For instance, the analysis presented in Fig. 4.8 shows that participants least interacted in locate tasks where multi-step interactions are mostly utilized, thus, reflecting interaction behavior and potential reasoning for the insignificant performance difference against static maps. The result suggests that users' interactivity is inversely proportional to the number of steps required to

perform the interaction.

4.7.3 Limitations & Future Work

Despite the interactive features significantly improving the correctness of the geospatial data exploration, the overall accuracy of visual estimation still requires further improvement, as shown in Fig. 4.3. We observed participants struggling in advanced exploratory tasks more than the simpler ones such as identify, and locate. Particularly in associate, categorize, and cluster types, the tasks where in-depth spatial correlation and comparison require, participants achieved relatively lower correctness scores spending significantly more time than the other type of tasks as shown in Fig. 4.4. Since the effect of interactive features is evident in the exploration visual analysis, there are plenty of scopes to implement advanced interactive features that facilitate these tasks. In our future research, we intend to formally survey the domain experts to outline the requirements for the interactive visual exploration of geospatial data. We plan to further measure the advantage of coordinated visualization including additional visual components in the environment. For the spatiotemporal exploration, we plan to implement a multi-layer contour on a single map instance and design the user interaction for exploratory purposes. We also plan to include experts in geospatial analysis in both quantitative and qualitative evaluations of the interactive geospatial visualization. While we leverage the temperature and precipitation intensity in this study, correspondingly as short and large spatial extent data, Hence, we can utilize the result for implementing user interaction that supports more advanced analysis tasks.

In this chapter, we report the usability of contour-based interactive geospatial visualization in exploratory analysis tasks. We implemented an interactive geospatial contour map and design a web-based interactive user study to quantitatively evaluate users' performance compared to the static geospatial map. We crowdsourced the study and recruited 62 participants from Amazon MTurk for our experiment. The

interactive features are scrutinized against seven exploratory analysis tasks using temperature and precipitation contour maps. The result suggests that the participants' correctness score improved with the interactive maps, however, needing additional time to complete the task compared to the static maps. We also analyzed participants' interaction logs to report and justify the usability of the interactive features in the exploratory geospatial tasks. Our findings can serve as guidelines for visualization researchers in the future to develop use case-specific interactive geovisualization for visually exploring geospatial data.

CHAPTER 5: VISUAL ANALYTICS APPROACH FOR LARGE-SCALE SPATIOTEMPORAL DATA

In this chapter, we introduce a novel visual analytics approach, DCPViz, to enable climate scientists to explore massive climate data interactively without requiring the upfront movement of massive data. Thus, climate scientists are afforded more effective approaches to support the identification of potential trends and patterns in climate projections and their subsequent impacts. We designed the DCPViz pipeline to fetch and extract NEX-DCP30 data with minimal data transfer from their public sources. We implemented DCPViz to demonstrate its scalability and scientific value and to evaluate its utility under three use cases based on different models and through domain expert feedback.

5.1 Introduction

In the United States (U.S.), the National Climate Assessment (NCA) [176] is the nation's leading resource for exploring and addressing adaptation and mitigation questions related to climate change across the various sectors of society. To explore and assess the potential impacts of future climates, the NCA uses statistically downscaled climate projections at a high spatial resolution of 10km or finer. Statistical downscaling (SD) is a correction technique for mitigating biases in important variables of interest (e.g., temperature and precipitation). This technique uses observations to estimate the bias-corrected variables appropriately and at a higher spatial resolution than the original model outputs.

Recognizing the importance and need for bias-corrected statistical downscaling to make reliable future climate projections (e.g., [175]), NASA's Earth eXchange project released the Downscaled Climate Projections (NEX-DCP30) at 800m resolution to support climate change research [21, 13]. The NEX-DCP30 provides climate projection data, including monthly averaged precipitation (pr), daily maximum tempera-

ture (*tasmax*), and daily minimum temperature (*tasmin*), for the contiguous U.S. for the 115 years between 1985 and 2099. The high-resolution spatiotemporal dataset is organized according to the Network Common Data Form (NetCDF) format [52] and publicly available on the Amazon Simple Storage Service (S3) [1].

There are, however, many challenges that limit the utility of the NEX-DCP30 and similar downscaled climate projection (DCP) data for climate research and relevant scientific analyses. First, due to the size of such data, it is often problematic for climate scientists to migrate data to local storage, which is often necessary for data analysis and exploration. For example, each NetCDF file contains 5 years of data over the contiguous U.S. at 800m resolution, which makes each file size approximately 2GB (or, 17TB for the complete archive) [115, 21]. Similarly, CMIP6 [211] is expected to produce more than 30 PB of simulation results through the Earth System Grid Federation. Second, existing tools for visualizing, analyzing, and exploring data like the NEX-DCP30 (e.g., Panoply [97], NCCV [13]) are limited. Existing tools normally depict only a few static plots with minimal support for interactive exploration and comparative analyses. Finally, even with prior knowledge of climate models and their associated data formats (e.g., NetCDF), it is a prohibitive and complex task to convert data into formats that are easily interpretable, understandable, and, thus, useful for decision-makers [59, 13]. [2].

In a traditional pipeline, (Fig. 5.1), analyzing massive climate data usually begins with users (e.g., climate scientists) searching and locating high-resolution climate data. Users, then, must manage local storage, prepare computational resources, and request data access. Once access is granted, users must download the (usually massive) data, convert the data into the desired format, analyze the data computationally, and, visualize the converted data for scientific analyses to create, for example, performance benchmarks based on the dataset. Due to the importance of high-resolution

[1]Data Source - https://nasanex.s3.amazonaws.com/NEX-DCP30/
[2]ESGF - https://esgf.llnl.gov/

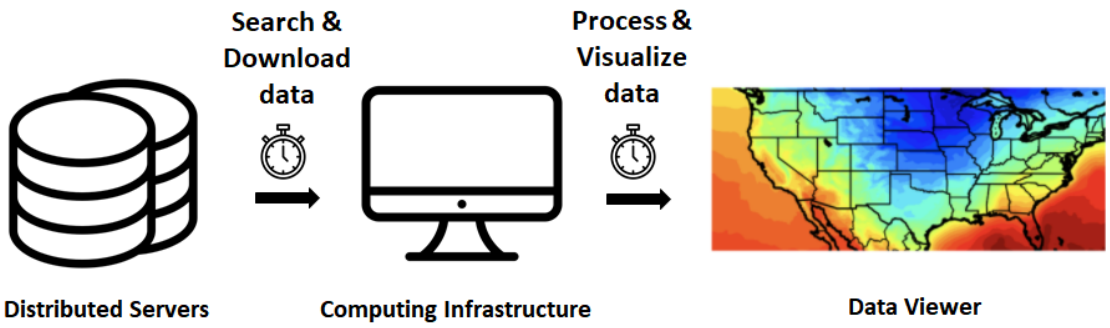


Figure 5.1: Traditional workflow for the climate scientists to explore the NEX-DCP30 data files. The climate scientists fetch the data files in a local machine from public data repositories (e.g., S3), manually filter by model name, variable, experiment ID, and year range, then, prepare the data and visualize it in the local machine using a NetCDF viewer.

Figure 5.2: DCPViz employs coordinated multi-views (CMV) to explore high-resolution statistically downscaled climate projections. It includes a regional view of (A) *Temporal Heatmap* - presents relative monthly-averaged projection of the climate variable (e.g., precipitation)(B) *Contour Matrix* - aligned with temporal heat map illustrates geo-coordinated projections, and (C) *Map view* - renders multi-layered interactive map for exploring geospatial contours.

climate data to the U.S. NCA [176] and other international climate assessments (e.g., the IPCC) [212], climate scientists and decision-makers require more innovative and interactive, visual, and cloud-based solutions to overcome the highlighted challenges and maximize the utility of climate data like the NEX-DCP30 to improve our understanding of future climate change at multiple scales (e.g., regional, national).

To address the challenges of analyzing and exploring shared high-resolution climate projection data like the NEX-DCP30, this paper introduces a novel interactive visual analytics approach, called DCPViz, that improves the efficiency and efficacy of climate

projection analyses. DCPViz provides data exploration and comparative analyses with better accessibility to shared data, and a streamlined environment for climate scientists and decision-makers to analyze climate projections data. To evaluate our approach, we demonstrated DCPViz to climate scientists, allowed the scientists to explore DCPViz (and associate climate data - i.e., NEX-DCP30), solicited expert feedback, and performed a qualitative evaluation of the system. Our findings revealed that key aspects of DCPViz - its interactivity, accessibility, and sense-making support - assisted the evaluators in climate analyses and improved overall task performance. The contributions of this paper are:

- We provide a novel cloud-based pipeline to extract high-resolution climate data (e.g., NEX-DCP30 data) for an interactive visual analytics system. The pipeline substantially reduces redundant downloads of large-scale data by moving analysis to the site of the data and migrating only the results of the analyses.
- We introduce a web-based interactive visual analytics tool, DCPViz, that helps scientists perform analysis, reasoning, and decision-making tasks with spatiotemporal climate projections.
- We present DCPViz use cases designed and recommended by climate scientists. We also present findings based on expert evaluations. These findings highlight the aspects of DCPViz that assist climate scientists to improve task performance.

5.2 Related Work

To situate the contributions of DCPViz, we summarize related works from two perspectives: i) visualization systems for earth science domains and ii) visualization techniques for relevant data types. Our literature review is further scoped to the challenges of analyzing massive, high-resolution DCP data.

5.2.1 Visualization Systems for Earth Science

Several visualizations and modeling platforms have been developed for climate data analysis in recent years. However, to our best knowledge, no previous platforms demonstrate either: i) a visual analytics pipeline with exploratory features for massive, high-resolution DCP data; or ii) sufficient interactive visualizations for exploratory and comparative analyses.

5.2.1.1 General Purpose Systems for Earth Science

Most general-purpose visualization systems demand data be organized according to a well-defined format prior to visualizing, exploring, and analyzing the data. In addition, visualization systems until recently were not compatible with cloud-based architectures. Analytics systems for large-scale data such as Community Data Analysis Tool (CDAT) [99, 28], Climate Engine [43], MATLAB [103], and NCAR [105], and GrADS [104] have been widely applied to the earth science domains. The goal of these platforms is to provide data mapping and time-series visualization features to users of geospatial data. Despite comprehensive analytical features, these systems do not offer integrated visualizations for analyzing spatiotemporal patterns.

5.2.1.2 Visualization Systems for Climate Change

Climate scientists work with diversified data that frequently demand visualizations and interactions specifically tailored for the analysis of massive, high-resolution spatiotemporal data. Hence, most of the special purpose systems have domain experts involved in the design, development, and evaluation process [213] according to task-specific requirements such as operational weather forecasting, climate change assessments, etc.

HI-RAMA [12] and VAPOR [34] are domain-specific systems to help climate scientists predict and explore the impact of potential climate change. NCCV [13] offers exploratory visualizations for NEX-DCP30 temperature and precipitation data. To

our best knowledge, NCCV is the only tool focused on NEX-DCP models, but it was developed based on Flash player which is no longer available. Similar visualization systems have been developed to facilitate deep exploration and comparative analysis of climate projections [214, 40, 39, 34]. But these systems limit analytical support to specific variables or models. NASA developed an online analysis platform, Giovanni [98], that limits its study of geophysical variables including precipitation and temperature, as well as, restricts users from loading preferred data and processing scripts.

Several platforms have been developed (e.g., MeteoInfo [198], WebGlobe [39], Panoply [97]) to analyze the data encoded in the NetCDF format, an abstraction for storing multidimensional data, and a popular data encoding format for climate variables [57, 39, 34]. These tools are sufficient to generate a geospatial view from NetCDF data but lack interactive features for exploring spatial patterns.

5.2.2 Interactive Visualization Techniques

Climate variables are often attributed to geolocation and timestamp information. As such, climate model evaluations utilize geospatial visualizations and time charts [213]. Geospatial visualizations commonly enable data exploration based on geographic location [76, 13]. Time charts are often used to identify trends over time. Information access, navigation, user interaction, and manipulation are some common affordances of interactive geospatial visualization systems. While visualization has a lengthy history with climatology, researchers continue to search for better visual approaches to explore complex environmental information in a manner that is both understandable and useful for a broader range of tasks [76]. As such, visual analytics systems often leverage other visual artifacts to support geospatial visualization. For example, area maps are used to plot the area of modeled impact [98, 34]. Parallel coordinates are utilized to analyze the association in multivariate climate data [42]. Histograms, bar plots, and hierarchical ensemble clustering are popularly leveraged

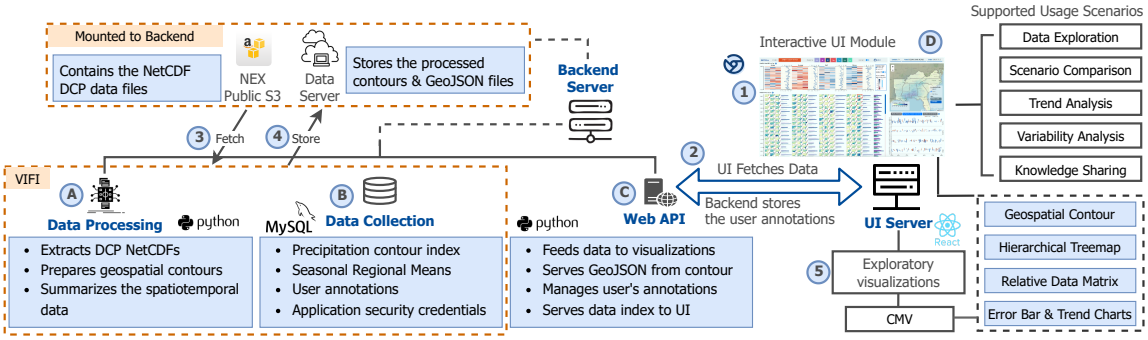


Figure 5.3: The proposed DCPViz pipeline incorporates 3 distributed servers for exploring spatiotemporal climate projections: 1) Data Server, 2) Backend Server, and 3) UI server. There are four major modules in our pipeline: A) *Data Processing Module* - performs transformation and analysis in a cloud environment (VIFI), B) *Data Collection Module* - stores the resulting indexed data, C) *Web API Module* - maintains the transactions between data and UI servers, and D) *Interactive UI* - a web-based module to allow users simultaneous access to the system.

in several systems to compare the climate models in different spatiotemporal granularity [48, 44]. However, these tools and systems provide insufficient support for coordinated multi-visualizations with interlinked user interactions, which are very useful for pattern detection and comparative analysis.

DCPViz, however, allows climate scientists to explore spatiotemporal patterns and covariability among projected climate variables using interactive spatiotemporal and geospatial visualization. The interactive features also allow the user to interact with its system pipeline to transform the data and link them to coordinated visualizations.

5.3 Design Method & Requirements

To design DCPViz, our team of visualization researchers and climate scientists met weekly for more than one year to define design requirements, discuss climate science workflows, and conduct several research and design interactions. We began our design integration using a traditional workflow (Fig. 5.1). Climate projection data are complex due to high dimensionality, multi-variability, and multi-resolution factors [52]. Exploratory analysis can benefit interactive transformation and filtering of complex

multidimensional data. However, such analyses require efficient information extraction from NetCDF data sources. Data extraction unfolds spatiotemporal information from NetCDF files and interprets data for interactive visualization. Visual analytics approaches enable the exploration of underlying patterns in different spatiotemporal granularity as compared to static visualizations. In addition, we reviewed related works and the goals of NEX for the NCA [21, 215]. These goals include: a) engaging scientists in addressing climate change, b) improving the scope of climate research and analysis, and c) facilitating better access to data and software tools to encourage collaboration and knowledge sharing. Based on the literature, we identified five primary design requirements:

R1 **To facilitate data preparation without the overhead of massive data downloads.** The system should extract spatiotemporal climate projections from NetCDF data files stored in a cloud environment efficiently and eliminate redundant downloads of large data files in order to run exploratory analyses.

R2 **To enable interactive transformation and filtering to multidimensional data.** The system should provide seamless access to interactive features for data transformation and filtering for exploratory and sense-making tasks. Leveraging a web-based protocol for data access will also support a NEX objective by enabling better data accessibility to climate scientists.

R3 **To enable climate scientists to perform visual exploratory analyses.** The system should offer a web-based interactive interface that allows climate scientists to access model projections for emission scenarios. The interface should provide climate scientists with visual components to perform analysis tasks, reason over analytical results, and identify latent insights.

R4 **To facilitate trend and covariance analysis in different spatiotemporal granularity.** The system should offer interactive visualizations to observe

underlying patterns and identify covariance among climate variables at multiple spatial (e.g., region, state, county) and temporal (e.g., annual, seasonal, monthly) granularities.

R5 **To enhance collaboration and sharing among climate scientists.** The system should serve as a platform for climate scientists where they can perform data exploration and analyze climate variables simultaneously. The system should facilitate further collaboration by enabling the sharing of key findings among climate scientists.

To satisfy the aforementioned requirements, we designed a novel visual analytics pipeline for DCPViz. This pipeline contains four modules: data processing, data collection, (RESTful) web API, and the UI, which provides interactive visual components. Fig. 5.3 illustrates the modules and process flows.

The Data Processing Module enables data transformation and analysis over distributed data by leveraging VIFI [216, 127]. This module runs on a remote server to access the NetCDF files from a mounted S3 bucket containing the DCP data (Fig. 5.3A and 5.3-3). These data are extracted from the cloud in an efficient matter to facilitate data preparation and reduce the overhead of massive data downloads prior to analysis **(R1)**.

The Data Collection Module stores the extracted data in a shared repository and obtains the data index to facilitate accessibility 'on demand' (Fig. 5.3B and 5.3-4). Each index is composed of a unique code for the model name, variable name, temporal identifier, and the DCP dataset name (e.g., *'NEX-DCP_CESM1-CAM5_pr_2021-03-01'*). Each index identifies its original NetCDF file. The shared repository stores the indices for the entire dataset (e.g., spatiotemporal precipitation, maximum and minimum temperatures). When the user focuses on a certain segment in the timeline or a spatial region, the interface passes a request to the API module to fetch the required data **(R2)**.

The Web API Module is a backend server that extracts a subset of the data from the remote site using specified filters (Fig. 5.3C). To eliminate the need for prerequisite knowledge about the climate model and data format, the system automatically pre-processes the data into the requested format. This is a standalone service that provides precipitation and temperature data to the UI module and associated analysis scripts **(R2)**.

The UI Module provides Coordinated Multi Views (CMV) to allow users to investigate climate change on a regional scale leveraging high-resolution NEX-DCP30 data [213]. This module provides not only conventional plots according to the traditional workflow but also an interactive visual interface (Fig. 5.3D). The interface supports exploratory analyses on future climate change over spatial and temporal domains **(R3)**. Moreover, the interface displays climate variability and a summary of overall seasonal changes between *pr* and *tasmax* and across the 7 NCA regions in the contiguous U.S. **(R4)**. Furthermore, DCPViz supports textual annotations in the map and matrix views to share key findings with other users on DCPViz **(R5)**. Section 5.4.3 describes the UI in detail.

We integrated these modules to build our proposed visual analytics system. The UI module interprets a sequence of events from the user and delegates them to the distributed modules to support requested functionalities. The system allows the user to foster efficient and collaborative research on regional climate change over the contiguous US without downloading and preprocessing data using local storage and without detailed knowledge of the data format and organization.

5.4 Visual Analytics Approach

In this section, we describe DCPViz and illustrate how it meets our design requirements. We begin by describing the NEX-DCP30 dataset that we leveraged to demonstrate the DCPViz UI. Next, we describe the data processing, data collection, and web API modules to present our novel cloud-based pipeline for transformation

and exploratory analysis of the high-resolution DCP data. We conclude by presenting UI Module, including its views and affordances for interaction.

5.4.1 The NEX-DCP30 Dataset

NEX is a collaborative platform to facilitate the analysis and forecast of climate data [21]. NEX projections provide several statistically downscaled datasets via a public S3 bucket. Among these datasets, NEX-DCP30 is the dataset with the highest spatial resolution of about 800m. Three variables are included in NEX-DCP30 (precipitation - ***pr*** **(**mm/day**)**, daily maximum temperature - ***tasmax*** (K), and daily minimum temperature - ***tasmin*** (K) later converted to (^oC)). They are derived from 36 climate models and observational data. We use two of these models for DCPViz: Community Atmospheric Model version 5 (CESM1-CAM5 [217]) and Goddard Institute for Space Studies Model E2 coupled with the Russell ocean model (GISS-E2-R [218]). NEX-DCP30 contains 115 years of monthly data for the contiguous U.S. Data from 1985 to 2005 are historical and data from 2006 to 2099 are projected under three RCP scenarios (2.6, 4.5, and 8.5).

5.4.2 Data Transformation & Analysis

We leverage VIFI, an open-source tool that enables the analysis of distributed fragmented data. VIFI enables data sharing by migrating analytics (often lightweight) to data locations rather than by migrating massive data [216, 127]. Users can perform a variety of analyses, specified as analytical workflows (Fig. 5.4), over distributed data stored at multiple locations. VIFI workflows are implemented for data-driven discoveries from distributed climate data [**?**]. We utilize VIFI to enable contour map generation, analyze climate variability and identify anomalous patterns from the RCP scenario projections (Fig 5.3A-B). Workflow execution begins with query generation based on user-selected parameters such as model name, climate variables, experiment ID, and year range. Once the backend server accesses the data corresponding to

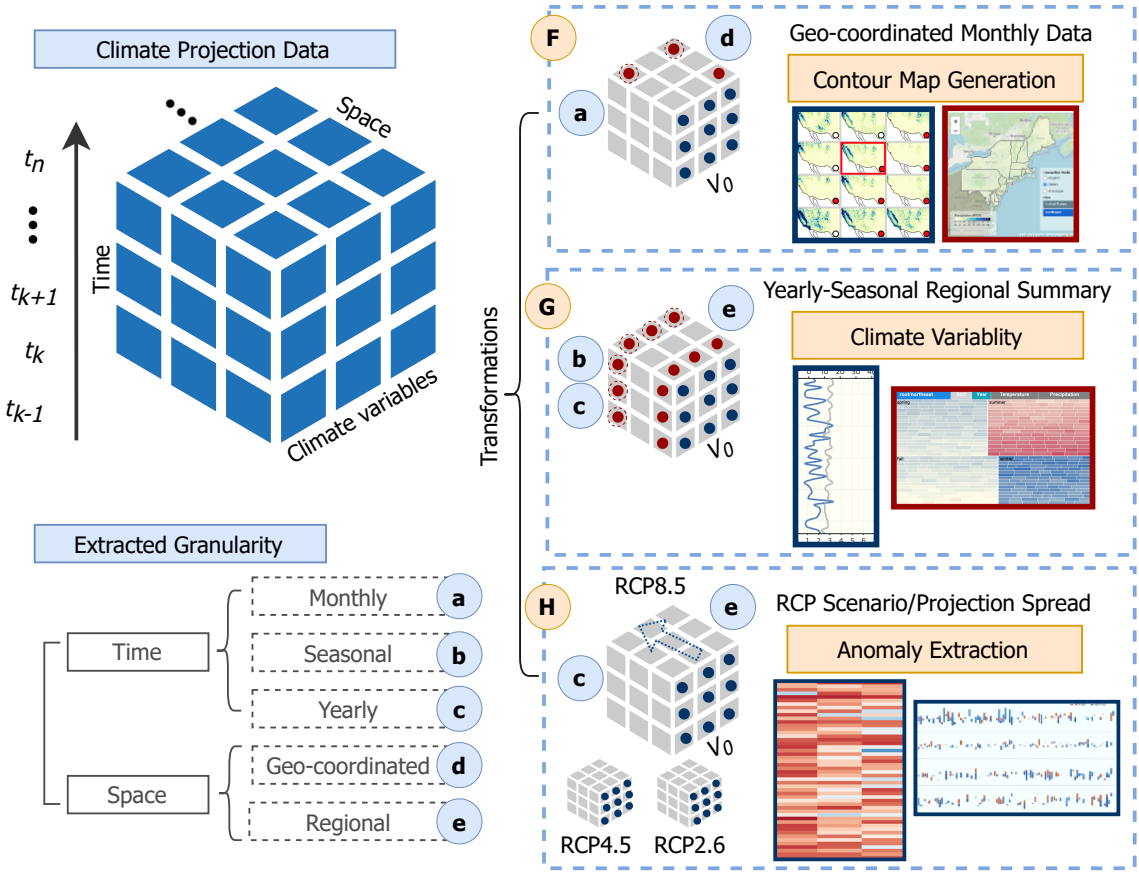


Figure 5.4: Spatiotemporal transformations of the projection data for exploratory analysis. The transformations illustrate the extracted hierarchical levels of spatial and temporal granularity.

specified parameters, the Data Processing Module (Fig. 5.3A) triggers the VIFI workflows to perform data extraction, transformation, and analysis. To demonstrate this process in DCPViz, we provide an illustration using *pr* and *tasmax* in the NEX-DCP30 models.

Contour Map Generation. We extract NetCDF files using VIFI data orchestration to retrieve geo-coordinated monthly averaged climate projections (e.g, *pr* and *tasmax*). The workflow extracts 60 projected snapshots (one for each month) from each NetCDF file that recorded the spatiotemporal climate variables. Next, the contour generation process (Fig. 5.4F) transforms the geo-coordinated data to the GeoJSON format. While extracting the data remotely, the collection module attaches an

index to the generated GeoJSON using the search parameters.

Spatiotemporal Granularity Extraction. We apply a series of transformations to geo-coordinated monthly projection data to extract the spatiotemporal granularity (Fig. 5.4). The 4th NCA report segments US territories in 10 regions [182], of which we have segmented the climate projections for 7 regions, including the Northeast, Southeast, Midwest, Southwest, Northwest, Northern Great Plains, and South Great Plains. The Alaskan, Hawaiian, and US Caribbean regions are excluded from the NEX-DCP30 data. We apply an NCA's region mask on the projected data to segment the temporal variable intensity into seven regions (Northeast, Southeast, Midwest, Southwest, Northwest, Northern Great Plains, and Southern Great Plains).[3] Then, we utilized the segmented data to calculate seasonal and yearly regional means for the projected variables. These data are further exploited to analyze and assess climate variability in the projected models. The data processing module passes the metadata, GeoJSON index, and extracted temporal data to the data collection module.

Extracting Anomalous Projections. Lastly, we derive measures to extract anomalous patterns from the climate projections. A previously calculated yearly mean is leveraged to measure the percentage change between the yearly mean and the selected decades of retrospective mean (e.g., 1985-2006), what we call *'Relative Intensity'*. Seasonal regional retrospective means are described as $RM(t, s, r)$ in the following:

$$RM(t, s, r) = \frac{1}{3\Delta t} * \sum_{y=t_0}^{t_1} \sum_{m=s}^{s+3} X_r(y, m) \tag{5.1}$$

where t represents a year range within retrospective timeline, s represents the season, and r represents the region. $X_r(y, m)$ provides the monthly-averaged regional mean for the climate variables. $RM(t, s)$ is leveraged in calculating the *'Relative Intensity'*,

[3]The 4th NCA report segments U.S. in 10 regions [182] but Alaskan, Hawaiian, and U.S. Caribbean are excluded in the NEX-DCP30 data

$RI(t, y, m, r)$ for pr described as follows:

$$RI(t, y, m, r) = \frac{\mid RM(t, \lfloor m/3 \rfloor, r) - X_r(y, m) \mid}{RM(t, \lfloor m/3 \rfloor, r)} \tag{5.2}$$

The intensity fluctuation is often projected within a wide range of variations, which can overlook anomalies at certain data points. For example, a 0.5 mm/day change in pr may not be noticeable in a snapshot compared to the overall range. However, this small change could be significant if the pr trends around 2 mm/day in that month/season. As such, $RI(t, y, m, r)$ can amplify anomalous patterns. The implementation of user-controlled thresholds and interactions facilitates this anomaly detection on the UI module as discussed in Section 5.4.4.

The **Data Collection Module** (Fig. 5.3B) maintains a repository of extracted and transformed data as well as analytical results. We store these data in a manner that allows for efficient filtering while exploring projections from the entire timeline. The extracted regional climate variables are summarized as monthly and yearly means and saved in another table to feed the summary visualizations. In addition, this module manages the user's credentials and annotations.

The **Web API Module** transforms collected data as per the required structure for the interactive visualizations before responding to the requests from the UI Module. All of these processes are executed remotely to limit data migration and the need for user knowledge of data format and organization.

5.4.3 DCPViz: Exploratory Visual Interface

The DCPViz interface presents interactive temporal and spatial visualizations for retrospective and projected climate data. The UI leverages CMV to enable rich user interaction. It consists of three main views: 1) a map view (Fig. 5.5C), 2) an adaptive spatiotemporal view (Fig. 5.5A-B), and 3) a reconfigurable summary panel (Fig. 5.6, 5.7, and 5.8). We implemented the interface for exploring the projection of historical

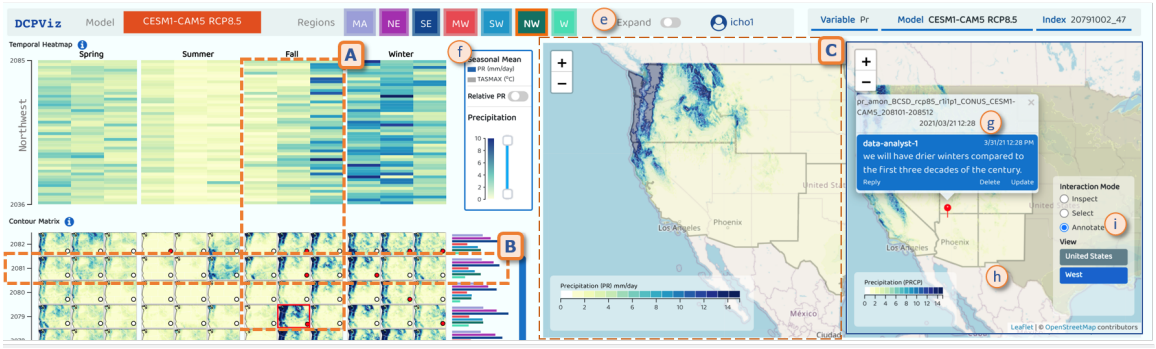


Figure 5.5: Overview of the DCPViz interface. (A) Temporal heatmap mapped with contour matrix presents a seasonal progression of geospatial intensity (*pr*) (top to bottom), whereas (B) regional bar chart provides yearly-averaged regional intensity (*pr*) followed by a monthly progression (left to right). (C) Map view enables annotating the geo-coordinated location in addition to interactive depictions of contour matrix snapshots.

pr and *tasmax* data across the U.S. The interface is developed using open-source web libraries.[4]

5.4.3.1 Map View

The map view shows a geo-coordinated contour that represents the monthly-averaged intensity across the U.S. regions. The view has three map layers: a base map, a contour map, and an area border map for the contiguous U.S. The base map shows different geographical features of the U.S. such as boundaries, rivers, and highways. The geospatial contour map is plotted as a variable intensity layer. The area border map shows the borders of the seven U.S. regions.

The map view provides three interaction modes (Fig. 5.5i): 1) inspection, 2) selection, and 3) annotation. The inspection mode allows users to find areas of similar intensity on the contour layer. Areas are grouped by the projected intensity levels, which range from 0-14 mm/day with a yellow-green-blue scheme. Hovering over a specific location opens a tooltip that shows additional information such as the inspected region and the monthly-averaged value at the selected snapshot. Hovering also strengthens the color intensity to let the user effortlessly differentiate the focused

[4]The interface is publicly available at https://esva.jpllab.net/

area from other areas (Fig. 5.2C). The selection mode allows users to indicate attention in a specific location. Downscaled climate data are projected at 800m resolution [115, 21]. So, concentrating attention on a specific region can provide essential detail that may not be easily observed from the default view. Users can also change the view hierarchy to the entire U.S., a specific region, or a specific state on the map.

The annotation mode allows users to share findings and observations on the map view. Once the user adds an annotation, the view shows a pin to share points of interest with others for further exploration (Fig. 5.5h). An additional marker is added to each thumbnail in the spatiotemporal view to indicate whether the annotation is added for that snapshot (Fig. 5.2B). This interaction also allows users to post annotations directly without marking a specific location on the map.

5.4.3.2 Spatiotemporal View

The spatiotemporal view presents thumbnails of geospatial contours with the corresponding temporal mean from 2036 to 2099. The entire dataset for the model contains 150 years of data [115] of which we extracted remotely the most recent 1380 (monthly average for 115 years) recorded/projected snapshots. Observations reveal that it can be difficult to identify trends or mark points of anomaly by traversing snapshots using only geospatial projections on the map. To address this, we developed the spatiotemporal view to render the geospatial data against a temporal axis. We add color-coded legends (Fig. 5.2f) with corresponding measuring units to denote *pr* and *tasmax* in the visualizations. This view has four visual components (Fig. 5.2A-B): a temporal heat map, a geo-spatiotemporal contour matrix, a climate variability time series, and a regional bar chart.

Temporal Heatmap. The temporal heatmap shows the projected mean *pr* sharing the identical axis with the temporal geospatial contour (Fig. 5.2C). We used two different measures to set the intensity for the heatmap: 1) mean *pr* for each snapshot (Fig. 5.5A); and, 2) relative *pr* (Fig. 5.2A) derived as $RI_{Pr}(t, y, m, r)$ in section 5.4.2.

Fig. 5.2e provides a control *'Relative PR'* to switch between these measures. We also add the option to filter this view by setting upper and lower limits of *pr*. A similar seasonal grouping is applied in the temporal geospatial thumbnail view. Fig. 5.5A shows an example of the Fall season where the cells in a row represent months in the corresponding year. The temporal heatmap helps users understand the progression of *pr* over time. The yellow-green-blue color scheme is used for monthly-averaged mean in the geospatial visualization [219]. In the *'Relative PR'* view, we adjust the color scheme to blue-pale yellow-red because it is popular for showing anomalies [219], which in our case denotes maximum increase to maximum decrease from blue to red.

Geo-spatiotemporal Contour Matrix. When capturing the contour for each geo-coordinated monthly snapshot, we store them as static images. These images are essential to show a preview to help users identify the point of interest. DCPViz presents the captured geospatial snapshot against a temporal axis. Therefore, the user obtains both the time information and suggested *pr* intensity before deciding to explore the contour in the map view. We maintain a uniform alignment between the temporal heatmap and geospatial contour matrix to help the user identify the corresponding contours from the heatmap cell. We also group the contour images by seasons (Fig. 5.2B) to help users perceive the seasonal *pr* trend. Hovering over thumbnails shows enlarged renderings of the contour with metadata for a detailed view.

Climate Variability Time-Series. The line charts represent a time series of mean *pr* (blue) and mean *tasmax* (gray) (Fig. 5.2A). *pr* and *tasmax* are subset data for the selected region, grouped by season and rendered for each season. Each data point in the line charts presents the yearly seasonal-average *pr* record and their associated change in the *tasmax*.

Regional Bar Chart. The regional bar presents the yearly seasonal-average *pr* grouped by region. The map view provides an indication of the intensity of *pr*

in different regions. However, the map view may not be adequate for the user to compare *pr* among the regions. Hence, the regional bar chart view is aligned with the geospatial contour in the temporal matrix, as shown in Fig. 5.5B, to show a precise comparison of *pr* among the regions. In addition, the bar representing the selected region is highlighted to amplify the seasonal comparison among regions when scrolling up or down. A yearly-averaged regional bar chart is also included to reveal yearly changes among the seven regions.

5.4.3.3 Summary Views

Climate projections are often large and complex, and thus difficult to visualize both in totality and with granularity. As such, we provide a reconfigurable summary view to encode and visualize the data. We considered three design factors for the summary view: the visualization purposes, the supporting analysis tasks, and the data type [220]. We leverage the multi-level spatiotemporal granularity extracted from the projected data during data transformation. The summary view contains three visualizations: a time-series visualization (Fig. 5.6), an RCP scenario comparison (Fig. 5.7), and a hierarchical treemap (Fig. 5.8A).

Time-Series Visualization. The time-series visualization shows the seasonal *pr* and *tasmax* mean from 1985 to 2098 (Fig. 5.6). It allows users to explore the trends of projected climate variables, such as *pr* and *tasmax*. It also provides support for direct search to help users find anomalies and patterns in the timeline. We found the time-series visualization to be a suitable candidate for identifying outliers and patterns across a large number of data points [221, 79].

RCP Scenario Comparison. A different design of a stacked bar chart is adopted for visual comparison and anomaly detection across RCP scenario projections (Fig. 5.7A). Each RCP scenario is distinguished by color, where overlapping areas are highlighted with strengthened intensity. This visualization presents a yearly-seasonal regional projection spread of the extreme RCP scenarios (e.g., RCP 4.5 and RCP

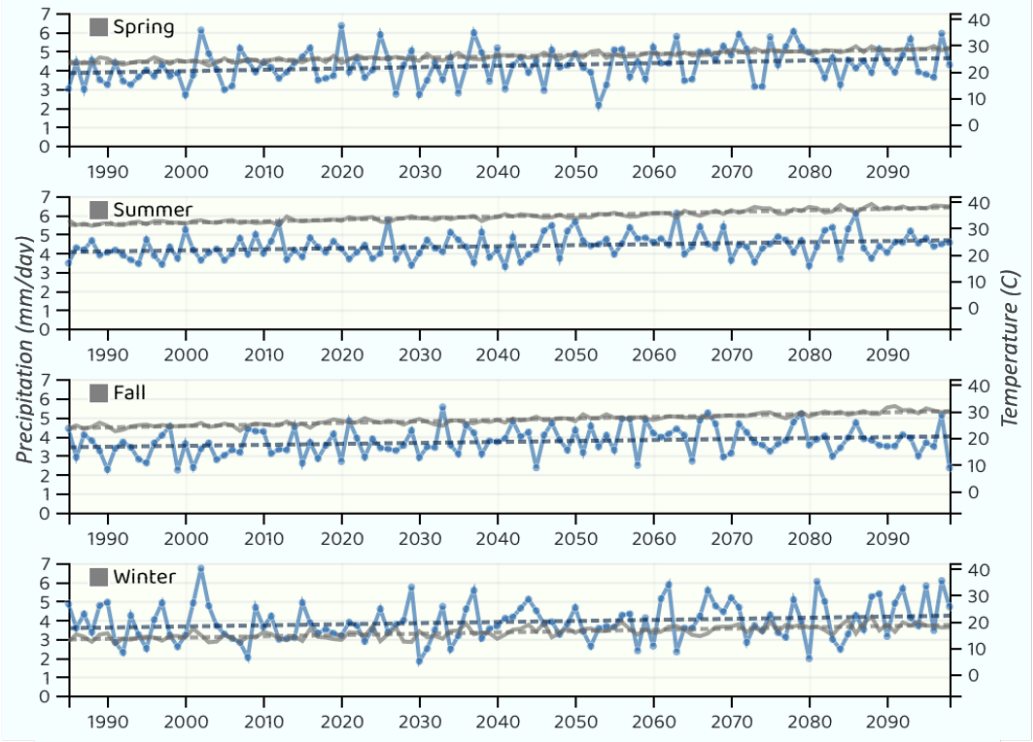


Figure 5.6: Time-series presents a summary view of yearly-averaged regional climate variability (e.g., projected *pr* and *tasmax* in the southeast region).

8.5 as compared to RCP 2.6). In Fig. 5.7B, we compare the scenarios for *tasmax* during the summer season in the western region according to CESM1-CAM5. Lower fluctuations among different RCP scenarios are observed in earlier decades in the 21st century. However, *tasmax* is projected to be significantly higher in the later years in the most extreme scenario.

Hierarchical Treemap. The treemap visualizes data with a nested structure for two selected climate variables, e.g., *pr* and *tasmax* (Fig. 5.8). The treemap is constructed maintaining the following hierarchy of data from the top: region, season, and year (Fig. 5.4). *pr* is denoted by the dimension of the cells whereas *tasmax* is denoted by the intensity using a blue-yellow-red color scheme for lowest to highest maximums. In the treemap, users can select an area of interest (Fig. 5.8B). The treemap presents an overview of the climate co-variability between variables to highlight, for example, which region is getting hotter & wetter or hotter & drier under a changing climate.

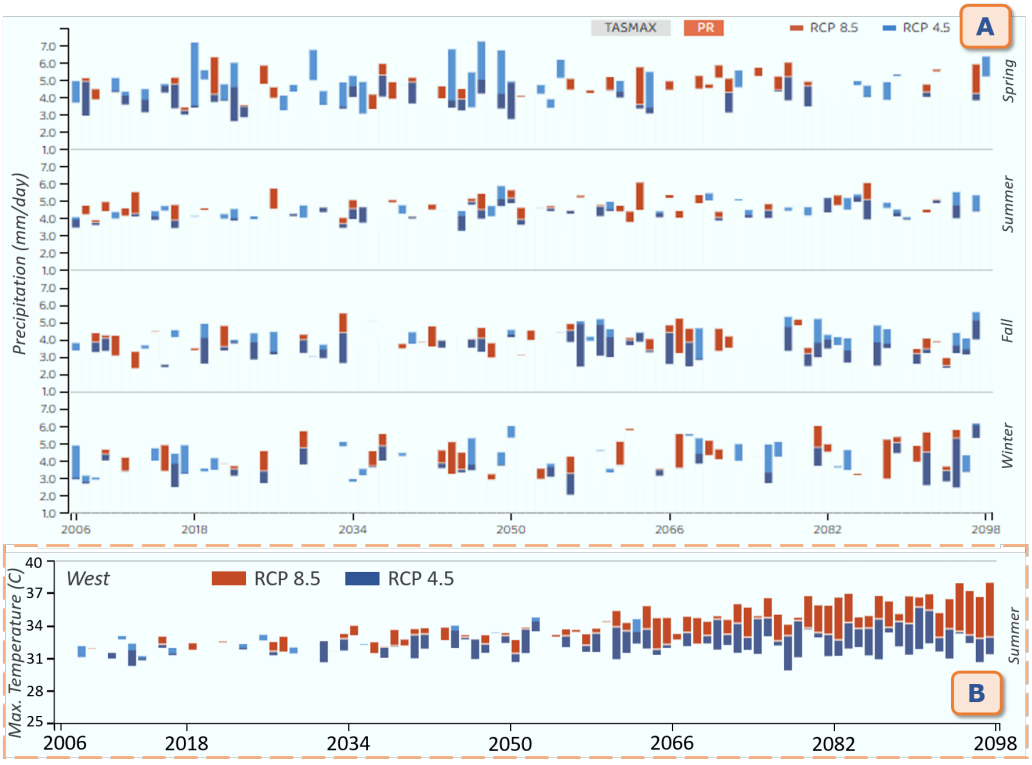


Figure 5.7: A) Comparative view of RCP scenarios to amplify patterns and anomalies in long-term climate projections. B) Comparing the *tasmax* projections in different RCP scenarios during summer seasons in the western region.

5.4.4 Interactions

We enable user interactions based on data and view manipulation capabilities to facilitate the human cognition process. These user interactions are categorized based on their support for different visual analysis tasks [94]. As each is used extensively, only representative illustrations are described.

Select. Users can select a region from the navigation panel and map view to switch the context for exploring region-specific data. We also enable select interactions to reconfigure the perspective view of summary visualizations among the time-series view, scenario comparison, and hierarchical treemap. This interaction often works as a precursor to other interaction techniques [96, 94] such as explore, connect, and filter.

Explore. When analyzing climate data, the explore interaction plays a vital role

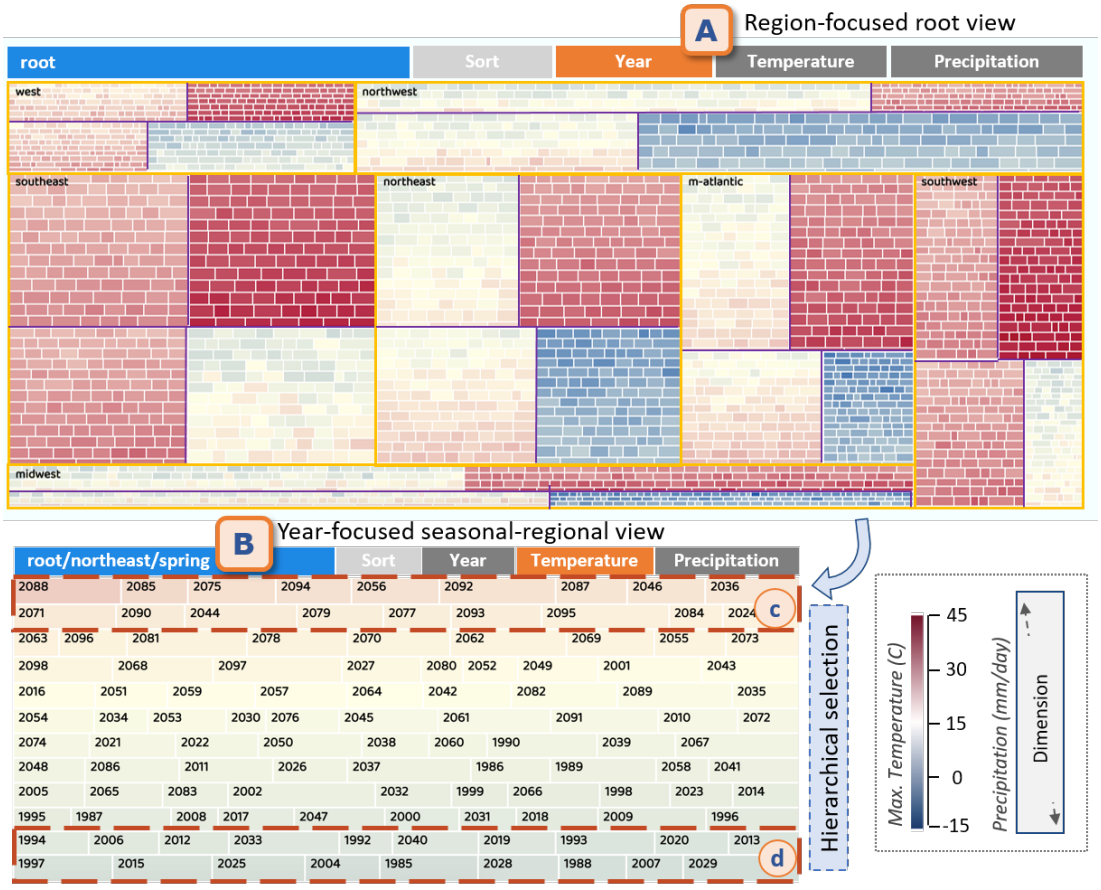


Figure 5.8: The hierarchical treemap supports investigating covariability between *pr* and *tasmax* across the 7 U.S. regions. It provides hierarchical selection to reconfigure the focus based on region, season, and year.

[222]. Often, explore interactions allow users to create additional views either by removing or overlapping a component [94] in an attempt to unfold underlying knowledge from multi-variate spatiotemporal data. In DCPViz, all visualizations provide additional spatial and temporal information in the form of a tooltip associated with each data point. Hovering over a thumbnail in the contour matrix (Fig. 5.5B) reveals an enlarged image of the thumbnail. This additional view also depicts timestamps, monthly-averaged intensity, and a static geospatial contour view. Left mouse clicks on a thumbnail (Fig. 5.2B) will present a dialog box for annotations as shown in Fig. 5.5g.

Filter. Users may filter based on spatial, temporal, and data value ranges. DCPViz will conditionally 'gray out' the heatmap cells that do not match the filter condition.

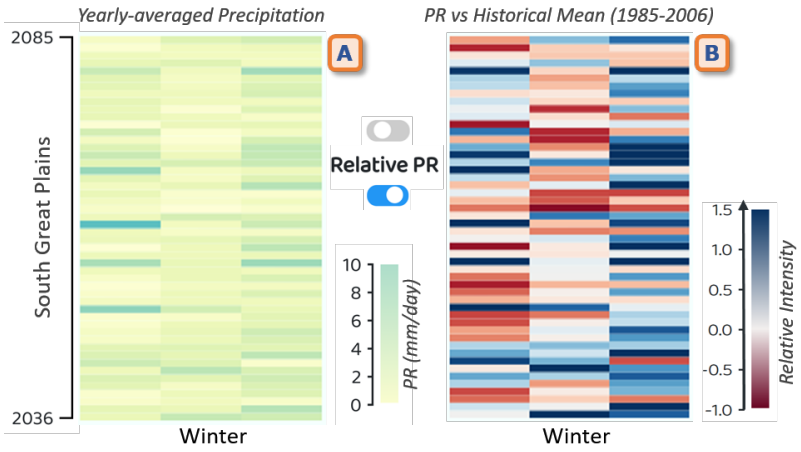


Figure 5.9: A scenario for the reconfigurable temporal heatmap of the Southern Great Plains in the winter season. The unnoticeable fluctuations in A) yearly-averaged intensity are strengthened in the (B) relative view by plotting the projections against a seasonal historical mean.

In addition, we add tiny red circles to the bottom-right of thumbnails in the spatiotemporal matrix to filter the projections that contain annotations from the user (Fig. 5.2f).

Abstract. This interaction allows users to elaborate on data from various levels of granularity, which is essential for visualizing large datasets. Fig. 5.8 demonstrates the abstract interaction, transitioning display from A to B when the user selects a region and B to C when the user selects a season. Geospatial visualization also provides abstract interaction where users can select regions and states for in-depth views, utilizing the high-resolution intensity in DCP.

Connect. The spatiotemporal, geospatial, and summary views are interlinked based on selected regions from the map view or the navigation panel. Each cell on the heatmap represents a corresponding thumbnail on the spatiotemporal matrix. The user can select a cell (Fig. 5.2g) to navigate among thumbnails (Fig. 5.2f). Selecting a thumbnail from the spatiotemporal matrix loads the geospatial *pr* contour into the map visualization connecting the interaction between these two views.

Reconfigure. This interaction supports rearranging visual representations of the

data. We provide reconfigure interactions to expand or shrink the coordinated views. The expanded view has additional visual components such as the climate variability trend and regional bar chart for each season. The temporal heatmap can be reconfigured to enable or disable the *'Relative PR'* view as shown in Fig. 5.9. Based on the user's preferences and analysis task, the interface allows users to reorder the summary visualizations described in Section 5.4.3.3. In addition, the hierarchical treemap provides sorting options to allow users to rearrange the tiles on the map. We also provide year, temperature, and precipitation as sorting options to inspect the same data from different perspectives.

5.5 Evaluation

Climate scientists provided a qualitative evaluation of DCPViz to determine its effectiveness at meeting its design requirements **(R1-R5)**, characterized the scientific value of climate data analyses, and demonstrated our research contributions. In this section, we describe the DCPViz evaluation metrics, the case studies utilized, and domain expert feedback.

5.5.1 Qualitative Evaluation Metrics

Evaluations of climate science visualizations span numerous qualitative and quantitative dimensions. For our evaluation, we focused on qualitative metrics that assess DCPViz based on its ability to reveal nuances of climate change, intricately visualize massive spatiotemporal data, and depict multivariate association factors for climate change [223]. To assess the dimensions of interlinked analysis tasks, communication, and decision-making in climate research, we used evaluation metrics from the following categories: visualization, interaction, and presentation.

Visualization. To evaluate the exploratory features of the visualizations, we assessed their ability to provide inexplicable insights to users from projected data. We wanted to determine the value our proposed visualizations added to the climate sci-

ence data analysis tasks. This question raised factors such as the time, complexity, and expertise [224] required to perceive a specific insight with and without DCPViz.

Interaction. Support for user interaction is central to the design of DCPViz. We focused our assessment on whether or not DCPViz support for interaction is sufficient to identify specific, relevant details within climate data. DCPViz utilizes CMVs. Thus, our qualitative assessments focused on the utility of filter, focus, and connect interactions for various exploratory tasks. We assessed the extent to which these interactions help users to exclude irrelevant parameters and values from the view and to discover the target data.

Presentation. We evaluated the visualizations based on their self-exploratory capabilities. To determine the quality of the presentation, we sought answers to the following three questions: 1) Are the colors, legends, and tooltips sufficient to understand the visualizations? 2) How easy is it to navigate among visualizations during analysis tasks?; and 3) Which visualizations are most useful for corresponding data and analyses? Based on these, we assessed whether domain experts are able to use the DCPViz without training.

5.5.2 Evaluation Procedure

We employed a survey to solicit feedback from domain experts on DCPViz. Seven climate scientists, who are experts in analyzing NEX-DCP30 data, participated in the survey. The survey contained three sections. First, participants were asked to complete the pre-questionnaire regarding their experience with climate data, analysis tasks, and the tools and data analysis libraries they utilize in their analysis. Next, participants watched a video that demonstrated DCPViz functions. Then, they interacted with the UI to explore NEX-DCP30 data. Finally, participants were asked to complete a post-questionnaire that inquires about: 1) Preference of the visualizations, 2) Features that support their analysis tasks, 3) Performance of the CMVs for explanatory capabilities and features.

5.5.3 Use Cases

We highlight the three use cases where DCPViz demonstrated the greatest value to climate scientists.

Comprehensive investigation of the spatiotemporal variability in regional precipitation (UC1). The temporal heatmap (Fig. 5.2A) revealed to domain experts future *pr* changes for each of the four seasons. For example, when the user selects the Southeast U.S. region (Fig. 5.5e) and enables *'Relative PR'* (Fig. 5.2e), the heat map cells in the winter season are mostly rendered in red colors (Fig. 5.5A). This essentially indicates a relatively drier winter later (2036-2085) compared to the first three decades of the century. The contour matrix (Fig. 5.5B) reveals that most of the region's *pr* occurs along the West Coast and Sierra Nevada. Users can move their cursors over the thumbnails to see the spatial patterns in *pr*. By clicking the thumbnail, the precipitation intensity is displayed on the map view with boundaries of the associated NCA region (Fig. 5.5C). The map view supports zooming to maximize the utility of the high spatial resolution (800m). Making annotations (Fig. 5.2g) facilitates a comparison of *pr* patterns between dry and wet months.

Analyses of long-term trends in temperature and precipitation under different emission scenarios (UC2). The summary view (Fig. 5.6) reveals in the climate data the overall warming trend of *tasmax* for all the seven NCA regions. The observed warming from 1985 to the present is well known, and the bias-corrected temperature from NEX-DCP30 is expected to represent the observed trend. Unlike the warming trends, however, observed *pr* trends in the climate data, as revealed by the relevant visualizations, appear not to be significant across the contiguous U.S.. By default, the line and bar charts represent CESM1-CAM5 model predictions under the RCP 8.5 emissions scenario, which represents high emissions of greenhouse gases without any mitigation efforts. Using DCPViz, domain experts were able to observe that across all seven NCA regions and four seasons, the associated warming trends

are higher than those for the RCP 4.5 and 2.6 scenarios. The plots indicate that *pr* changes are more variable than *tasmax* and *tasmin* in all three scenarios. Also, the *pr* predicted by CESM1-CAM5 does not exhibit consistent sensitivity to the emission scenarios across the seven regions (Fig. 5.7A).

Covariability of seasonal maximum temperature and precipitation (UC3). The hierarchical treemap is one of the unique visual components leveraged in DCPViz that is rarely utilized for analyzing spatiotemporal data. The hierarchical treemap (Fig. 5.8) provides scope to observe co-variability between the climate variables. The cell dimension represents mean *pr* while color intensity scales temperature. The hierarchical treemap also helps domain experts answer key scientific questions related to climate projections - e.g., Are warmer and more humid conditions expected during the latter decades of the 21st century in the fall season of the Western regions? What about spring in the Northeast region?

5.5.4 Domain Expert Feedback

The domain experts who participated in the survey possessed on average 5 years of experience in climate science research. In the pre-questionnaire, they mentioned that climate data processing and visualization is a moderately to highly difficult task due to the volume and complexity of high-dimensional structure. The domain experts also expressed their desire to have interactive geospatial and time-series views, in contrast to traditional static plots (e.g., by Python or Matlab).

Visualizations. We asked domain experts which visualizations they find useful on the interface and for what purposes. The domain experts mentioned that interactive geospatial visualization (Fig. 5.5C) is worthwhile to have and appreciated the annotation feature as it allows them to share the observations with other scientists. Additionally, the inspect user interactions helped them identify the spatial pattern in each snapshot **(UC1)**. The spatiotemporal view (Fig. 5.5A) was also identified as effective by four experts to understand the temporal progression along with the spatial

distribution. The heatmap helped the domain experts quickly inspect the data as the heatmap contains the summary for the whole timeline. The temporal contour matrix aided the domain experts with a quick view of the spatial patterns and their changes over time. We also learned from the pre-questionnaire that the time-series chart is the most common visualization for climate scientists for trend analysis. Along with the seasonal and regional time charts, the domain experts identified the treemap as a less common yet useful visualization for analyzing comparative patterns for multiple variables.

Interactions. The experts' feedback suggests that the reconfigure and explore interactions are essential for their analysis. One expert stated that the reconfigurable view for summary visualization is convenient for sensemaking long-term patterns from different aggregation perspectives **(UC2, UC3)**. The exploratory features such as inspecting the map view from the contour thumbnail (Fig. 5.5B) and switching the multi-views based on region selection helped the domain experts to identify quickly interesting facts or insights from the data **(R3)**. Moreover, the tooltips were well appreciated as they provide comprehensive data inspection capabilities and additional context in the visual analysis. We have received positive feedback on the data inspection capability in general for all the visualizations on the interface.

Presentation. We asked the participants how convenient it is for them to discover the target information. The participants responded stating the interface is convenient for them to interact. Among them, one participant labeled the interface as easy to use after the initial learning process. The domain experts were able to discover their target information from the interactive visualizations while one expert stated the interface became intuitive after the initial learning process. Another participant quoted, *"This tool is great for making quick inspections of the data as not only does the heatmap provide a summary for multiple timesteps, but the map interface lets me quickly see what the spatial patterns look like for each timestep, helping me get a better*

idea at how spatial patterns change with time." Moreover, the experts mentioned that the treemap is a good candidate to illustrate the covariability between *pr* and *tasmax* **(UC3)**. We also learned that the use of information icons, color codes, and legends were deemed useful by them. The domain experts also left us a few suggestions such as - 1) to include annotation for other visualizations (beyond the map), 2) to increase the use of animation in visualizing the projections, and 3) to attempt to simplify the UI.

The domain experts also provided some suggestions to improve the interface. One expert pointed out that the interface is a bit complex with a handful of content for a relatively small window. Two other experts suggested that we add more textual details about the visualizations and reduce the use of acronyms. They also desired more pervasive support for the annotation feature in other visualizations (beyond the map view). In addition, they suggested opportunities for animation - e.g., to show the change in projected *pr* for a selected time period - which could benefit their analysis.

Based on the majority of the feedback, we understand DCPViz benefits domain experts as they perform exploratory analyses of climate data **(UC1)**. Moreover, the use cases suggest that DCPViz is also useful in examining long-term trends within the NCA regions **(UC2, UC3)**.

5.6 Discussion and Limitations

We proposed a cloud-based visual analytics approach, DCPViz, for making sense of DCP datasets that benefits exploratory analysis by climate scientists. We described specific challenges and identified important design requirements for systems that support the analysis of massive high-resolution spatiotemporal data for climate change research. We hypothesized that our proposed requirements would enhance the traditional analysis workflow and support climate scientists in sense-making and decision-making processes. We described the four modules that comprise DCPViz (data processing, data collection, web API, and UI) to satisfy the design requirements

of the DCPViz pipeline (Fig. 5.3). In addition, we demonstrated how DCPViz can enhance transitional analysis workflows and support climate scientists' sensemaking and decision-making.

5.6.1 System Scalability

Initially, we prepared a computational script for query generation and extracting NetCDF files. Data extraction itself for geospatial contour was a relatively straightforward task since climate scientists apply the same extraction in their traditional workflow on a regular basis. We identified that the required time window for loading NetCDF and processing the GeoJSON is massive. Thus, our initial challenge was to reduce the data loading and extraction time. We distributed the computation in multiple threads at once to significantly reduce the data processing time. Since the data loading requires a substantial amount of primary memory, we had to find a proper balance of thread count and memory availability in our backend server. We parameterized these features so that we can tweak the processing module when the backend is moved to a different server. The complex behavior of the climate projection data raised the requirement to introduce a data collection module in the system. A database system supported optimized searching capability for a large amount of metadata and index for generated GeoJSONs. We demonstrated DCPViz with two NEX-DCP30 models (CESMI-CAM5 and GISS-E2-R) in 3 RCP scenarios. The web-based UI provides simultaneous access to climate scientists to visually explore these extracted projection data. However, climate models are continuously updated to provide high-fidelity simulations by changing parameters. With computational resources becoming more easily accessible, the climate simulations with finer spatial and temporal resolutions (e.g., NEX-GDDP [225]), daily projections are becoming increasingly available. To incorporate other DCP models and RCP scenarios, we modularized the data processing and collection modules. This allows the scientists to trigger the processing module to extract more model data and store it in the collection module

(R1).

Moreover, we understand that climate scientists' exploratory analyses demand more analytical workflows than we presented in Section 5.4.2. The DCPViz pipeline allows climate scientists to incorporate additional analytical workflows leveraging VIFI [44] in the backend. The results from the workflows are required to satisfy one of the transformations presented in Fig. 5.4 to visualize the DCPViz UI. Furthermore, to perform a comparative visual analysis between different data and scenarios, the user can simultaneously open several DCPViz interfaces for different DCP data or RCP scenarios. For example, the user can open one DCPViz interface with the CESM1-CAM5 model and another with the GISS-E2-R model for multi-model scenario comparison.

5.6.2 Interactive visualizations

Geospatial visualization and time series charts are popular visualization techniques for climate scientists [213] which we identified from the experts' feedback as well. Our design study shows that the select and explore interactions in the geospatial view allow the users to understand the spatial patterns of precipitation from different perspectives such as seasonal, regional, and state-wide. To foster the environment for collaboration in the earth science community, we introduced the annotation feature on the spatial data so that the scientist can share their observations. In addition, we introduced a novel multi-view spatiotemporal visualization to inspect a spatial pattern against a temporal axis. To address the challenge of browsing the large-scale climate projection data, we employed CMVs to support seamless navigation to the entire timeline of the projection model (Fig. 5.5A). Furthermore, we provided a summary panel on the interface providing trends and patterns between precipitation and maximum temperature from three different perspectives to explore the long-term model projection **(R3)**. Based on the user study feedback for **(UC1)**, we learn that our design choices for the geospatial visualization are sublime for exploratory analysis of geospatial data. The idea of annotating spatial data is appreciated by domain

experts. They suggested supporting it with annotation for other visualizations as well **(R4)**. They also commended the swiftness of the spatiotemporal views to inspect the prominent aspect of the data at a glance. The summary visualizations presented on the interface (Fig. 5.5D, Fig. 5.8, and Fig. 5.6) also found to be useful by the experts from the Expert's feedback evaluation. The hierarchical and sorting interaction for the treemap is particularly helpful for the experts to analyze covariability between maximum temperature and precipitation **(UC3)**.

5.6.3 Limitations

To support navigating and exploring the entire model projection, we have added exploratory and explanatory visualizations in the same window. As convenient as it is to gain insight promptly, the interface looks a bit complicated and it requires an initial learning curve to use the interactions. To perform a comparative analysis between projection models and scenarios, the user needs to open two DCPViz interfaces with the participant models or scenarios. In the future, we aim to support visual analysis of multi-model comparisons in DCPViz.

DCPViz enables the exploration of large volumes of data by creating visual components. We learned from the user study that system performance is affected by client-side network status and computational capacity. Despite the clear benefits of climate science analyses, system performance improvement will be part of our future work. Implementing combinable tabs [226] and adopting the progressive visualization pipeline [227] can ease users' analytical vision while the processing is running in the background. Employing a server-side tiled map technique for the map view can reduce the rendering time for geospatial contours to enhance the overall performance.

CHAPTER 6: RESEARCH GUIDELINES

In this dissertation, we have conducted a comprehensive review of the literature, identified the research scope, established research questions, and designed an exploratory approach to visual analytics for distributed spatiotemporal data. Based on the insights gained from this research, we outlined a set of guidelines for the visualization researchers to pursue further progress in the visual analytics for large-scale spatiotemporal data, as described below:

Hierarchical clustering for visualizing spatiotemporal granularities. In our data type classification, we found a majority of climate data falls into a multidimensional type due to the high dimensionality, multi-variability, and multi-resolution factors in climate data. In contrast, tree-type datasets are not much utilized in climate VA. However, we understand interpreting spatiotemporal data to tree data type holds enormous potential to observe patterns and identify correlations among climate variables through several spatial (e.g., region, state, and county) and temporal granularities (e.g., annual, seasonal, and monthly). Consequently, hierarchical clustering [48] and treemap [228, 113] visualization techniques are not extensively utilized in spatiotemporal climate data exploration and have the potential to be a new research direction for climate science visualizations.

Growing trend of integrated visualizations in spatiotemporal exploration. In recent years, integrated visualization techniques have become widely utilized as tools to visualize spatiotemporal climate data. Integrated visualizations are essential for the exploratory analysis of climate variables because of their usability and existing and potential future new interactive features. Visualizations that enable scientists to perceive salient data features quickly and efficiently are key. For example, distributing the spatial and temporal dimensions across multiple visual components allows users to inspect the data by focusing on individual dimensions. Consequently, such a capability not only helps scientists to identify underlying trends and patterns

from the data but also helps them to maximize cognitive efficiency. According to our review, CMV has become a trending approach for the interactive exploration of large-scale spatiotemporal climate data. However, significant challenges remain in terms of choosing relevant visualization techniques, display formation, and interactive features for exploratory visual analysis. Roberts [91] reflected on the view generation, interaction, and manipulation aspect to achieve state-of-the-art in CMV.

Underexplored immersive visual environments. Climate scientists often use point cloud data from observations and numerical models which can be analyzed in a fully immersive environment [229] applying Virtual and Augmented Reality technologies. Such an immersive analytic environment would provide better spatial perception to climate scientists for exploratory spatial analysis with other climate variables and a collaborative environment to discuss and share their findings with other scientists and researchers.

Interactive shepherding of exploratory models. Shepherd user interactions, discussed in 2.3.3, are underutilized in climate science VA. This entails a shortcoming of these systems in facilitating user-driven exploratory analysis. Systems such as UV-CDAT [28] and MeteoInfo [179] comprise active participation in building the exploratory analysis model. However, these systems are heavily code dependent and often lack interactive features (Task requirement **R10**). This unfolds a research opportunity to enable interactive user-driven exploratory analysis from the VA interface (Task requirement **R3**).

Case-specific approaches over general-purpose VA. We grouped the spatiotemporal VA systems for climate science into general-purpose and case-specific systems. While the general purpose systems [230, 28, 43] are well-commended for diverse usability and rich visualization library, these systems often lack the domain or use-case-specific custom visualizations and user interactions to support analysis tasks or demand high levels of programming expertise from the user. In addition,

we reviewed VA systems developed to serve specific analysis tasks or data sources [12, 24, 37, 113]. In contrast to the general purpose systems, these systems are more oriented to employ a use-case-specific analytics pipeline, visualization techniques, and user interaction. Climate scientists and visualization researchers usually engage in a joint collaboration to develop case-specific systems to satisfy the analysis requirements. Currently, a significant number of VA systems developed for climate science focus on specific use cases.

Surge of web-based VA systems. As climate scientists work with diversified datasets to understand climate change and its social impact and derive effective decision-making for adaptation, they often obtain data from disparate sources to perform distributed decision support and sensemaking analysis [231]. To conduct such analyses, they commonly transfer the data from distributed sources to their local workstations, apply data processing and transformations on the workstations, then generate visual illustrations. To mediate this inefficient workflow, a web-based VA approach can address these challenges of accessing disparate remote data sources, performing case-specific analyses on high-performance clusters, and providing an interactive interface for exploratory spatiotemporal visualizations [232, 113]. With advancements in modern web technology, VA systems are widely built on the web-based platform in recent years.

Low pervasion among climate scientists. A comprehensive review of several analytical reports and surveys [233, 27], as well as, our discussion with the atmospheric and environmental scientists identify that most climate scientists do not leverage state-of-the-art VA solutions to conduct their exploratory analyses. This gap between VA capabilities and climate science advancement must be bridged. Traditional approaches that incidentally rely on visualizations are becoming increasingly insufficient to match the scope of the climate science community's scientific challenges. Much of the information contained in the ever-growing observational and modeling

datasets remains untapped. For example, two-dimensional plots, that are central to publications and anchor scientific paper development are highly reductive. Training of climate scientists and collaboration with VA domain experts will help bridge these persistent gaps.

Unexploited potential of progressive visualization. There are growing numbers of observational platforms with increased data rates, growing numbers of climate models that exhibit increasing complexity, and a growing sophistication of experiments undertaken with climate models [234]. This rapid growth of data also challenges VA systems to support on-demand exploratory capabilities over large volumes of data. Conventional methods in VA systems often leverage background computation based on user queries [87]. They, then, display the computed result at once upon completion. Zgraggen et al. [227] labeled such a method as a 'blocking approach' where the user's analytical visions are blocked until the computation is fully completed. To address this challenge, climate research can benefit from the potential for progressive visualizations where data are incrementally processed in smaller chunks as the analytics systems interactively update the interface, working from an approximation that is refined over time [227, 25]. As the computations take place in chunks, shepherd interaction [94] can be utilized to build the exploratory analysis model. We have not found any progressive VA systems to explore spatiotemporal climate data. This indicates another opportunity for visualization researchers to develop progressive systems in collaboration with climate scientists.

CHAPTER 7: CONCLUSION & FUTURE WORK

This dissertation identifies the research questions for adopting a cloud-based architecture for distributed exploratory spatiotemporal analysis and correspondingly introduces three major components for interactive exploratory visual analytics for distributed spatiotemporal data. These components include a visual analytic framework presenting a data mining architecture, a contour-based exploratory geospatial visualization, and a visual analytics approach for large-scale distributed spatiotemporal data.

In chapter 3, we presented a visual analytics framework (VAF) for distributed data analysis systems (DAS) to mediate users' direct interaction with the distributed servers, provide access control from the application layer, and enable the exploratory visual analysis of results. To demonstrate the benefit of our proposed framework, we developed workflows for two use cases from earth science and SHBE research domains, working with respective domain experts. While we understand the potential of our VAF in distributed data analysis, we have several takeaways for future directions. Our future work will focus on complying with more distributed systems and developing a generic API service to deploy at DAS sites. Moreover, we aim to provide a more convenient interface for configuration management and perceive the user's domain knowledge to provide interactive visualization recommendations to explore the resulting data.

In chapter 4, we report the usability of contour-based interactive geospatial visualization in exploratory analysis tasks. We implemented an interactive geospatial contour map and design a web-based interactive user study to quantitatively evaluate users' performance compared to the static geospatial map. We crowdsourced the study and recruited 62 participants from Amazon MTurk for our experiment. The interactive features are scrutinized against seven exploratory analysis tasks using temperature and precipitation contour maps. The result suggests that the participants'

correctness score improved with the interactive maps, however, needing additional time to complete the task compared to the static maps. We also analyzed participants' interaction logs to report and justify the usability of the interactive features in the exploratory geospatial tasks. Our findings can serve as guidelines for visualization researchers in the future to develop use case-specific interactive visualization for visually exploring geospatial data.

In chapter 5, we introduced DCPViz, a novel visual analytics approach to analyze and explore large-scale distributed spatiotemporal data where we leveraged NASA's downscaled climate projections, the NEX-DCP30 data. To demonstrate the visual analytics interface, we leveraged the NEX-DCP30 data. We presented three use cases provided by climate scientists while interacting with the DCPViz interface. Experts' feedback is utilized to evaluate the proposed visual analytics approach. While the feedback we received from the experts is promising, we identified several future directions for further research. Our future work will encompass a data mining architecture that is fully integrated with a distributed framework to optimize the data transformation and analysis time. In addition, we will utilize the extracted data from multiple DCP models to assess the climate projection uncertainty. To evaluate the model projections for each climate variable, we plan to add observational datasets, more analytical workflows, comparative visualizations, and interactive features on the interface. Finally, we plan to conduct a more extensive user evaluation by gathering climate scientists in focus groups to evaluate in greater detail the usability and efficiency of DCPViz.

REFERENCES

[1] B. Shneiderman, "The eyes have it: a task by data type taxonomy for information visualizations," in *Proceedings 1996 IEEE Symposium on Visual Languages*, pp. 336–343, Sep. 1996.

[2] J. He, H. Chen, Y. Chen, X. Tang, and Y. Zou, "Variable-based spatiotemporal trajectory data visualization illustrated," *IEEE Access*, vol. 7, pp. 143646–143672, 2019.

[3] C. Tominski, G. Andrienko, N. Andrienko, S. Bleisch, S. I. Fabrikant, E. Mayr, S. Miksch, M. Pohl, and A. Skupin, "Toward flexible visual analytics augmented through smooth display transitions," *Visual Informatics*, vol. 5, pp. 28–38, 9 2021.

[4] N. Andrienko and G. Andrienko, "A visual analytics framework for spatio-temporal analysis and modelling," *Data Mining and Knowledge Discovery*, vol. 27, no. 1, pp. 55–83, 2013.

[5] W. Von Engelhardt, J. Zimmermann, and J. Zimmerman, *Theory of earth science*. CUP Archive, 1988.

[6] S. S. Board, N. R. Council, *et al.*, *Earth science and applications from space: national imperatives for the next decade and beyond*. National Academies Press, 2007.

[7] G. J. Borradaile, *Statistics of earth science data: their distribution in time, space and orientation*. Springer Science & Business Media, 2003.

[8] N. Andrienko, G. Andrienko, and P. Gatalsky, "Exploratory spatio-temporal visualization: An analytical review," *Journal of Visual Languages and Computing*, vol. 14, pp. 503–541, 2003.

[9] M. M. Alam, L. Torgo, and A. Bifet, "A survey on spatio-temporal data analytics systems," 2021.

[10] P. Rosen, A. Seth, B. Mills, A. Ginsburg, J. Kamenetzky, J. Kern, C. R. Johnson, and B. Wang, "Using contour trees in the analysis and visualization of radio astronomy data cubes," 4 2017.

[11] A. Comrie, K.-S. Wang, S.-C. Hsu, A. Moraghan, P. Harris, Q. Pang, A. Pińska, C.-C. Chiang, R. Simmonds, T.-H. Chang, *et al.*, "Carta: Cube analysis and rendering tool for astronomy," *Astrophysics Source Code Library*, pp. ascl–2103, 2021.

[12] J. H. McLean, S. B. Cleveland, M. Lucas, R. Longman, T. W. Giambelluca, J. Leigh, and G. A. Jacobs, "The hawai'i rainfall analysis and mapping application (hi-rama): Decision support and data visualization for statewide rainfall data," pp. 239–245, Association for Computing Machinery, 7 2020.

[13] J. R. Alder and S. W. Hostetler, "Web based visualization of large climate data sets," *Environmental Modelling Software*, vol. 68, pp. 175–180, 2015.

[14] Z. Li, Q. Huang, Y. Jiang, and F. Hu, "Sovas: a scalable online visual analytic system for big climate data analysis," *International Journal of Geographical Information Science*, vol. 34, no. 6, pp. 1188–1209, 2020.

[15] F. Kamw, S. Al-Dohuki, Y. Zhao, T. Eynon, D. Sheets, J. Yang, X. Ye, and W. Chen, "Urban structure accessibility modeling and visualization for joint spatiotemporal constraints," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 1, pp. 104–116, 2020.

[16] D. Sha, X. Miao, H. Lan, K. Stewart, S. Ruan, Y. Tian, Y. Tian, and C. Yang, "Spatiotemporal analysis of medical resource deficiencies in the us under covid-19 pandemic," *PloS one*, vol. 15, no. 10, p. e0240348, 2020.

[17] G. A. Meehl, W. M. Washington, J. M. Arblaster, A. Hu, H. Teng, J. E. Kay, A. Gettelman, D. M. Lawrence, B. M. Sanderson, and W. G. Strand, "Climate change projections in cesm1 (cam5) compared to ccsm4," *Journal of Climate*, vol. 26, no. 17, pp. 6287–6308, 2013.

[18] Y. Liu, A. R. Ganguly, and J. Dy, "Climate downscaling using ynet: A deep convolutional network with skip connections and fusion," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, (New York, NY, USA), p. 3145â3153, Association for Computing Machinery, 2020.

[19] R. S. Khan and M. A. E. Bhuiyan, "Artificial intelligence-based techniques for rainfall estimation integrating multisource precipitation datasets," *Atmosphere*, vol. 12, no. 10, 2021.

[20] M.-J. Kraak and D. E. van de Vlag, "Understanding spatiotemporal patterns: Visual ordering of space and time," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 42, no. 2, pp. 153–161, 2007.

[21] R. R. Nemani, B. L. Thrasher, W. Wang, T. J. Lee, F. S. Melton, J. L. Dungan, and A. Michaelis, "Nasa earth exchange (nex) supporting analyses for national climate assessments," vol. 2015, p. GC21Eâ04, 2015.

[22] A. Talukder, M. Elshambakey, S. Wadkar, H. Lee, L. Cinquini, S. Schlueter, I. Cho, W. Dou, and D. J. Crichton, "Vifi: Virtual information fabric infrastructure for data-driven discoveries from distributed earth science data," in *IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/ IOP/SCI)*, pp. 1–8, Aug 2017.

[23] C. A. L. Pahins, N. Ferreira, and J. L. Comba, "Real-time exploration of large spatiotemporal datasets based on order statistics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 11, pp. 3314–3326, 2020.

[24] M. Kalo, X. Zhou, L. Li, W. Tong, and R. Piltner, "Chapter 8 - sensing air quality: Spatiotemporal interpolation and visualization of real-time air pollution data for the contiguous united states," in *Spatiotemporal Analysis of Air Pollution and Its Application in Public Health* (L. Li, X. Zhou, and W. Tong, eds.), pp. 169–196, Elsevier, 2020.

[25] U. U. Turuncoglu, "Toward modular in situ visualization in earth system models: the regional modeling system regesm 1.1," *Geoscientific Model Development*, vol. 12, no. 1, pp. 233–259, 2019.

[26] K.-L. Ma, "In situ visualization at extreme scale: Challenges and opportunities," *IEEE Computer Graphics and Applications*, vol. 29, no. 6, pp. 14–19, 2009.

[27] S. Afzal, M. M. Hittawe, S. Ghani, T. Jamil, O. Knio, M. Hadwiger, and I. Hoteit, "The state of the art in visual analysis approaches for ocean and atmospheric datasets," in *Computer Graphics Forum*, vol. 38, pp. 881–907, Wiley Online Library, 2019.

[28] E. Santos, J. Poco, Y. Wei, S. Liu, B. Cook, D. N. Williams, and C. T. Silva, "Uv-cdat: Analyzing climate datasets from a user's perspective," *Computing in Science & Engineering*, vol. 15, no. 1, pp. 94–103, 2013.

[29] R. MacIejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert, "A visual analytics approach to understanding spatiotemporal hotspots," vol. 16, pp. 205–220, 3 2010.

[30] W. Chang, M. L. Stein, J. Wang, V. R. Kotamarthi, and E. J. Moyer, "Changes in spatiotemporal precipitation patterns in changing climate conditions," *Journal of Climate*, vol. 29, no. 23, pp. 8355 – 8376, 2016.

[31] H. Mohammadi, M. R. Delavar, M. A. Sharifi, and M. D. Pirooz, "Spatiotemporal visualization of tsunami waves using kml on google earth," vol. 42, pp. 1291–1299, International Society for Photogrammetry and Remote Sensing, 9 2017.

[32] C. Helbig, H.-S. Bauer, K. Rink, V. Wulfmeyer, M. Frank, and O. Kolditz, "Concept and workflow for 3d visualization of atmospheric data in a virtual reality environment for analytical approaches," *Environmental earth sciences*, vol. 72, no. 10, pp. 3767–3780, 2014.

[33] Y. Zhang, K. Sung, and J. Fridley, "A web framework for interactive data visualization system," *MSCSSE Capstone Rep. Comput. Softw. Syst. Univ. Wash. Bothell*, p. 84, 2019.

[34] S. Li, S. Jaroszynski, S. Pearse, L. Orf, and J. Clyne, "Vapor: A visualization package tailored to analyze simulation data in earth system science," *Atmosphere*, vol. 10, p. 488, 2019.

[35] J. Sukharev, C. Wang, K.-L. Ma, and A. T. Wittenberg, "Correlation study of time-varying multivariate climate data sets," in *2009 IEEE Pacific Visualization Symposium*, pp. 161–168, IEEE, April 2009.

[36] F. Wang, W. Li, S. Wang, and C. R. Johnson, "Association rules-based multivariate analysis and visualization of spatiotemporal climate data," *ISPRS International Journal of Geo-Information*, vol. 7, no. 7, 2018.

[37] C. Li, G. Baciu, Y. Wang, J. Chen, and C. Wang, "Ddlvis: Real-time visual query of spatiotemporal data distribution via density dictionary learning," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021.

[38] M. Maskey, R. Ramachandran, I. Gurung, M. Ramasubramanian, B. Freitag, A. Kaulfus, G. Priftis, D. Bollinger, R. Mestre, and D. da Silva, "Employing deep learning to enable visual exploration of earth science events," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2248–2251, Sep. 2020.

[39] A. Sharma, S. M. A. Zaidi, V. Chandola, M. R. Allen, and B. L. Bhaduri, "Webglobe - a cloud-based geospatial analysis framework for interacting with climate data," pp. 42–46, Association for Computing Machinery, Inc, 11 2018.

[40] J. Herring, M. S. VanDyke, R. G. Cummins, and F. Melton, "Communicating local climate risks online through an interactive data visualization," *Environmental Communication*, vol. 11, pp. 90–105, 1 2017.

[41] J. Johansson, T. Opach, E. Glaas, T.-S. Neset, C. Navarra, B.-O. LinnÃ©r, and J. K. RÃžd, "Visadapt: A visualization tool to support climate change adaptation," *IEEE Computer Graphics and Applications*, vol. 37, no. 2, pp. 54–65, 2017.

[42] J. Poco, A. Dasgupta, Y. Wei, W. Hargrove, C. Schwalm, R. Cook, E. Bertini, and C. Silva, "Similarityexplorer: A visual inter-comparison tool for multifaceted climate data," *Computer Graphics Forum*, vol. 33, no. 3, pp. 341–350, 2014.

[43] J. L. Huntington, K. C. Hegewisch, B. Daudert, C. G. Morton, J. T. Abatzoglou, D. J. McEvoy, and T. Erickson, "Climate engine: Cloud computing and visualization of climate and remote sensing data for advanced natural resource monitoring and process understanding," *Bulletin of the American Meteorological Society*, vol. 98, pp. 2397–2409, 11 2017.

[44] A.-A.-R. Nayeem, M. Elshambakey, T. Dobbs, H. Lee, D. Crichton, Y. Zhu, C. Chokwitthaya, W. J. Tolone, and I. Cho, "A visual analytics framework for

distributed data analysis systems," in *2021 IEEE International Conference on Big Data (Big Data)*, pp. 229–240, Dec 2021.

[45] S. Hadlak, C. Tominski, H.-J. Schulz, and H. Schumann, "Visualization of attributed hierarchical structures in a spatiotemporal context," *International Journal of Geographical Information Science*, vol. 24, no. 10, pp. 1497–1513, 2010.

[46] G. M. H. Zahan, D. Mondal, and C. Gutwin, "Contour line stylization to visualize multivariate information," 2021.

[47] T. Eaglin, I. Cho, and W. Ribarsky, "Space-time kernel density estimation for real-time interactive visual analytics," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[48] C. P. Kappe, M. Böttinger, and H. Leitte, "Analysis of decadal climate predictions with user-guided hierarchical ensemble clustering," in *Computer Graphics Forum*, vol. 38, pp. 505–515, 2019.

[49] L. Liu, M. Mirzargar, R. M. Kirby, R. Whitaker, and D. H. House, "Visualizing time-specific hurricane predictions, with uncertainty, from storm path ensembles," in *Computer Graphics Forum*, vol. 34, pp. 371–380, 2015.

[50] S. Lu, R. M. Li, W. C. Tjhi, K. K. Lee, L. Wang, X. Li, and D. Ma, "A framework for cloud-based large-scale data analytics and visualization: Case study on multiscale climate data," in *2011 IEEE Third International Conference on Cloud Computing Technology and Science*, pp. 618–622, Nov 2011.

[51] S. Kim, S. Jeong, I. Woo, Y. Jang, R. Maciejewski, and D. S. Ebert, "Data flow analysis and visualization for spatiotemporal statistical data without trajectory information," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, pp. 1287–1300, 3 2018.

[52] E. Edward Hartnett and R. Rew, "Experience with an enhanced netcdf data model and interface for scientific data access," in *24th Conference on IIPS*, 2008.

[53] R. O. Obe and L. S. Hsu, *PostgreSQL: Up and Running: a Practical Guide to the Advanced Open Source Database*. " O'Reilly Media, Inc.", 2017.

[54] R. Obe and L. Hsu, *PostGIS in action*. Simon and Schuster, 2021.

[55] M. D. Da Silva and H. L. Tavares, *Redis Essentials*. Packt Publishing Ltd, 2015.

[56] K. Chodorow, *MongoDB: the definitive guide: powerful and scalable data storage*. " O'Reilly Media, Inc.", 2013.

[57] R. Rew and G. Davis, "Netcdf: an interface for scientific data access," *IEEE computer graphics and applications*, vol. 10, pp. 76–82, 1990.

[58] P. Scarponi, G. Coro, and P. Pagano, "A collection of aquamaps native layers in netcdf format," *Data in Brief*, vol. 17, pp. 292–296, 4 2018.

[59] E. Davis, C. S. Zender, D. K. Arctur, K. O'Brien, A. Jelenak, D. Santek, M. J. Dixon, T. L. Whiteaker, and K. Yang, "Netcdf-cf: supporting earth system science with data access, analysis, and visualization," *AGUFM*, vol. 2017, p. IN33Câ0137, 2017.

[60] B. Eaton, J. Gregory, B. Drach, K. Taylor, S. Hankin, J. Caron, R. Signell, P. Bentley, G. Rappa, H. Höck, *et al.*, "Netcdf climate and forecast (cf) metadata conventions," 2003.

[61] J. C. Biard, J. Yu, M. Hedley, S. J. D. Cox, A. Leadbetter, N. J. Car, K. A. Druken, S. Nativi, and E. Davis, "Linking netcdf data with the semantic web-enhancing data discovery across domains," *AGUFM*, vol. 2016, p. IN23Aâ1758, 2016.

[62] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, pp. 1–10, 2010.

[63] C. A. Steed, J. R. Goodall, J. Chae, and A. Trofimov, "Crossvis: A visual analytics system for exploring heterogeneous multivariate data with applications to materials and climate sciences," *Graphics and Visual Computing*, vol. 3, p. 200013, 2020.

[64] N. Cao, C. Lin, Q. Zhu, Y. R. Lin, X. Teng, and X. Wen, "Voila: Visual anomaly detection and monitoring with streaming spatiotemporal data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, pp. 23–33, 1 2018.

[65] E. J. Pebesma, K. de Jong, and D. Briggs, "Interactive visualization of uncertain spatial and spatioâtemporal data under different scenarios: an air quality example," *International Journal of Geographical Information Science*, vol. 21, no. 5, pp. 515–527, 2007.

[66] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. Moorhead, "Noodles: A tool for visualization of numerical weather model ensemble uncertainty," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, pp. 1421–1430, 2010.

[67] Q. Shu, H. Guo, J. Liang, L. Che, J. Liu, and X. Yuan, "Ensemblegraph: Interactive visual analysis of spatiotemporal behaviors in ensemble simulation data," in *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 56–63, April 2016.

[68] D. Keim, "Information visualization and visual data mining," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 1–8, 2002.

[69] J. Li, K. Zhang, and Z.-P. Meng, "Vismate: Interactive visual analysis of station-based observation data on climate changes," in *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 133–142, IEEE, Oct 2014.

[70] E. Mayr, G. Schreder, S. Salisu, and F. Windhager, "Integrated visualization of space and time: A distributed cognition perspective," Dec 2018.

[71] S. Castruccio, M. G. Genton, and Y. Sun, "Visualizing spatiotemporal models with virtual reality: from fully immersive environments to applications in stereoscopic view," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 182, no. 2, pp. 379–387, 2019.

[72] S. Papers, M. Hlawitschka, T. Weinkauf, A. Thudt, D. Baur, and S. Carpendale, "Visits: A spatiotemporal visualization of location histories," 2013.

[73] N. Andrienko, G. Andrienko, E. Camossi, C. Claramunt, J. M. Cordero Garcia, G. Fuchs, M. Hadzagic, A.-L. Jousselme, C. Ray, D. Scarlatti, and G. Vouros, "Visual exploration of movement and event data with interactive time masks," *Visual Informatics*, vol. 1, no. 1, pp. 25–39, 2017.

[74] F. Ferstl, M. Kanzler, M. Rautenhaus, and R. Westermann, "Time-hierarchical clustering and visualization of weather forecast ensembles," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, pp. 831–840, Jan 2017.

[75] T. Nocke, T. Sterzel, M. Böttinger, M. Wrobel, *et al.*, "Visualization of climate and climate change data: An overview," *Digital earth summit on geoinformatics*, pp. 226–232, 2008.

[76] W. Cartwright, J. Crampton, G. Gartner, S. Miller, K. Mitchell, E. Siekierska, and J. Wood, "Geospatial information visualization user interface issues," *Cartography and Geographic Information Science*, vol. 28, pp. 45–60, 2000.

[77] M. Rautenhaus, M. Bottinger, S. Siemen, R. Hoffman, R. M. Kirby, M. Mirzargar, N. Rober, and R. Westermann, "Visualization in meteorology - a survey of techniques and tools for data analysis tasks," 12 2018.

[78] T. Nocke, M. Flechsig, and U. Bohm, "Visual exploration and evaluation of climate-related simulation data," in *2007 Winter Simulation Conference*, pp. 703–711, IEEE, 2007.

[79] T. Nocke, H. Schumann, and U. Böhm, "Methods for the visualization of clustered climate data," *Comput. Stat.*, vol. 19, p. 75â94, feb 2004.

[80] O. Schroth, E. Pond, S. Muir-Owen, C. Campbell, and S. R. Sheppard, "Tools for the understanding of spatio-temporal climate scenarios in local planning: Kimberley (bc) case study," *Zurich: Swiss National Sciences Foundation*, 2009.

[81] G. A. Schmidt, M. Kelley, L. Nazarenko, R. Ruedy, G. L. Russell, I. Aleinov, M. Bauer, S. E. Bauer, M. K. Bhat, R. Bleck, *et al.*, "Configuration and assessment of the giss modele2 contributions to the cmip5 archive," *Journal of Advances in Modeling Earth Systems*, vol. 6, no. 1, pp. 141–184, 2014.

[82] M.-J. Kraak, "The space-time cube revisited from a geovisualization perspective," in *Proc. 21st International Cartographic Conference*, pp. 1988–1996, Citeseer, 2003.

[83] M. Kraak, "Timelines, temporal resolution, temporal zoom and time geography," in *ICC 2005 : Proceedings of the 22nd international cartographic conference*, (New Zealand), International Cartographic Association, 2005. 22nd International Cartographic Conference, ICC 2005 : Mapping approaches into a changing world, ICC ; Conference date: 09-07-2005 Through 16-07-2005.

[84] C. Tominski, P. Schulze-Wollgast, and H. Schumann, "3d information visualization for time dependent data on maps," in *Ninth International Conference on Information Visualisation (IV'05)*, pp. 175–181, July 2005.

[85] N. Andrienko and G. Andrienko, "Spatio-temporal visual analytics: a vision for 2020s," *Journal of Spatial Information Science*, no. 20, pp. 87–95, 2020.

[86] N. Röber and J. F. Engels, "In-situ processing in climate science," in *High Performance Computing: ISC High Performance 2019 International Workshops, Frankfurt, Germany, June 16-20, 2019, Revised Selected Papers*, (Berlin, Heidelberg), p. 612â622, Springer-Verlag, 2019.

[87] K.-L. Ma, "In situ visualization at extreme scale: Challenges and opportunities," *IEEE Computer Graphics and Applications*, vol. 29, no. 6, pp. 14–19, 2009.

[88] P. S. Quinan and M. Meyer, "Visually comparing weather features in forecasts," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, pp. 389–398, Jan 2016.

[89] M. Sakr, G. Andrienko, T. Behr, N. Andrienko, R. H. Güting, and C. Hurter, "Exploring spatiotemporal patterns by integrating visual analytics with a moving objects database system," in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, (New York, NY, USA), p. 505â508, Association for Computing Machinery, 2011.

[90] Y. Q. Wang, "Meteoinfo: Gis software for meteorological data visualization and analysis," *Meteorological Applications*, vol. 21, no. 2, pp. 360–368, 2014.

[91] J. C. Roberts, "State of the art: Coordinated amp; multiple views in exploratory visualization," in *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*, pp. 61–71, July 2007.

[92] G. Andrienko and N. Andrienko, "Coordinated multiple views: a critical view," in *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*, pp. 72–74, IEEE, 2007.

[93] G. Convertino, J. Chen, B. Yost, Y.-S. Ryu, and C. North, "Exploring context switching and cognition in dual-view coordinated visualizations," in *Proceedings International Conference on Coordinated and Multiple Views in Exploratory Visualization - CMV 2003* -, pp. 55–62, July 2003.

[94] Y. Lu, R. Garcia, B. Hansen, M. Gleicher, and R. Maciejewski, "The state-of-the-art in predictive visual analytics," in *Computer Graphics Forum*, vol. 36, pp. 539–562, Wiley Online Library, 2017.

[95] E. Dimara and C. Perin, "What is interaction for data visualization?," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 119–129, 2019.

[96] J. S. Yi, Y. ah Kang, J. Stasko, and J. A. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *IEEE transactions on visualization and computer graphics*, vol. 13, no. 6, pp. 1224–1231, 2007.

[97] J. G. Acker, G. T. Alcott, M. Ventura, J. C. Wei, and D. J. Meyer, "The Advantages of Synergy - Quantitative Earth Science Data Visualization and Analysis with Giovanni, Panoply, and Excel," in *AGU Fall Meeting Abstracts*, vol. 2018, pp. IN21B–35, Dec. 2018.

[98] J. G. Acker and G. Leptoukh, "Online Analysis Enhances Use of NASA Earth Science Data," *EOS Transactions*, vol. 88, pp. 14–17, Jan. 2007.

[99] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. Johhson, "Visualization of uncertainty and ensemble data: Exploration of climate modeling and weather forecast data with integrated visus-cdat systems," in *Journal of Physics: Conference Series*, vol. 180, p. 012089, IOP Publishing, 2009.

[100] D. Cashman, S. R. Humayoun, F. Heimerl, K. Park, S. Das, J. Thompson, B. Saket, A. Mosca, J. Stasko, A. Endert, *et al.*, "A user-based visual analytics workflow for exploratory model analysis," in *Computer Graphics Forum*, vol. 38, pp. 185–199, Wiley Online Library, 2019.

[101] W. Cui, "Visual analytics: A comprehensive overview," *IEEE Access*, vol. 7, pp. 81555–81573, 2019.

[102] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim, "Knowledge generation model for visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1604–1613, 2014.

[103] D. Hanselman and B. Littlefield, *Mastering MATLAB: a comprehensive tutorial and reference*. Prentice-Hall, Inc., 1996.

[104] F. Berman, A. Chien, K. Cooper, J. Dongarra, I. Foster, D. Gannon, L. Johnsson, K. Kennedy, C. Kesselman, J. Mellor-Crumme, D. Reed, L. Torczon, and R. Wolski, "The grads project: Software support for high-level grid application development," *The International Journal of High Performance Computing Applications*, vol. 15, no. 4, pp. 327–344, 2001.

[105] D. Brown, R. Brownrigg, M. Haley, and W. Huang, "Ncar command language (ncl)," 2012.

[106] H. Lee, A. Goodman, L. McGibbney, D. E. Waliser, J. Kim, P. C. Loikith, P. B. Gibson, and E. C. Massoud, "Regional climate model evaluation system powered by apache open climate workbench v1.3.0: an enabling tool for facilitating regional climate studies," *Geoscientific Model Development*, vol. 11, no. 11, pp. 4435–4449, 2018.

[107] R. J. Sandusky, "Computational provenance: Dataone and implications for cultural heritage institutions," in *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3266–3271, IEEE, 2016.

[108] J. P. Cohn, "Dataone opens doors to scientists across disciplines," 2012.

[109] D. Medvedev, G. Lemson, and M. Rippin, "Sciserver compute: Bringing analysis close to the data," in *Proceedings of the 28th international conference on scientific and statistical database management*, pp. 1–4, 2016.

[110] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Würthwein, *et al.*, "The open science grid," in *Journal of Physics: Conference Series*, vol. 78, p. 012057, IOP Publishing, 2007.

[111] T. Nocke, T. STERZEL, M. BÃ¶ttinger, and M. Wrobel, "Visualization of climate and climate change data: An overview," *in Ehlers et al. (Eds.) Digital Earth Summit on Geoinformatics 2008: Tools for Global Change Research (ISDE'08), Wichmann, Heidelberg, pp. 226-232, 2008*, 01 2008.

[112] J. Kehrer, F. LadstÃ€dter, P. Muigg, H. Doleisch, A. Steiner, and H. Hauser, "Hypothesis generation in climate research with interactive visual data exploration," 2008.

[113] A.-A.-R. Nayeem, H. Lee, D. Han, M. Elshambakey, W. J. Tolone, T. Dobbs, D. Crichton, and I. Cho, "Dcpviz: A visual analytics approach for downscaled climate projections," pp. 291–300, 2022.

[114] A.-A.-R. Nayeem, I. Segovia-Dominguez, H. Lee, D. Han, Y. Chen, Z. Zhen, Y. Gel, and I. Cho, "Learning on health fairness and environmental justice via interactive visualization," in *2022 IEEE International Conference on Big Data (Big Data)*, pp. 784–791, 2022.

[115] H. Lee, A. Goodman, and K. Gorski, "The big climate data pipeline (bcdp): A tool to facilitate server-side processing of nasa-nex data on amazon cloud," *AGUFM*, vol. 2019, p. IN12Aâ03, 2019.

[116] J. W. Lee and S. Y. Hong, "Potential for added value to downscaled climate extremes over korea by increased resolution of a regional climate model," *Theoretical and Applied Climatology*, vol. 117, no. 3-4, pp. 667–677, 2014.

[117] N. Röber, M. Böttinger, and B. Stevens, "Visualization of climate science simulation data," *IEEE Computer Graphics and Applications*, vol. 41, no. 1, pp. 42–48, 2021.

[118] R. Fuchs and H. Hauser, "Visualization of multi-variate scientific data," *Computer Graphics Forum*, vol. 28, pp. 1670–1690, 2009.

[119] C. Tominski, J. F. Donges, and T. Nocke, "Information visualization in climate research," in *2011 15th International Conference on Information Visualisation*, pp. 298–305, IEEE, 2011.

[120] M. D. Gerst, M. A. Kenney, A. E. Baer, A. Speciale, J. F. Wolfinger, J. Gottschalck, S. Handel, M. Rosencrans, and D. Dewitt, "Using visualization science to improve expert and public understanding of probabilistic temperature and precipitation outlooks," *Weather, Climate, and Society*, vol. 12, no. 1, pp. 117–133, 2020.

[121] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of international conference on intelligence analysis*, vol. 5, pp. 2–4, McLean, VA, USA, 2005.

[122] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim, "Knowledge generation model for visual analytics," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1604–1613, 2014.

[123] A. Talukder, M. Elshambakey, S. Wadkar, H. Lee, L. Cinquini, S. Schlueter, I. Cho, W. Dou, and D. J. Crichton, "Vifi: Virtual information fabric infrastructure for data-driven discoveries from distributed earth science data," in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pp. 1–8, IEEE, 2017.

[124] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr, "Xsede: Accelerating scientific discovery," *Computing in Science Engineering*, vol. 16, pp. 62–74, Sep. 2014.

[125] R. J. Sandusky, "Computational provenance: Dataone and implications for cultural heritage institutions," in *IEEE International Conference on Big Data (Big Data)*, pp. 3266–3271, Dec 2016.

[126] S. Shahand, M. M. Jaghoori, A. Benabdelkader, J. L. Font-Calvo, J. Huguet, M. W. Caan, A. H. van Kampen, and S. D. Olabarriaga, *Computational Neuroscience Gateway: A Science Gateway Based on the WS-PGRADE/gUSE*, pp. 139–149. Cham: Springer International Publishing, 2014.

[127] M. Elshambakey, M. Khalefa, W. J. Tolone, S. D. Bhattacharjee, H. Lee, L. Cinquini, S. Schlueter, I. Cho, W. Dou, and D. J. Crichton, "Towards a distributed infrastructure for data-driven discoveries & analysis," in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 4738–4740, IEEE, 2017.

[128] C. Chokwitthaya, Y. Zhu, R. Dibiano, and S. Mukhopadhyay, "Combining context-aware design-specific data and building performance models to improve building performance predictions during design," *Automation in construction*, vol. 107, p. 102917, 2019.

[129] O. T. Karaguzel, M. Elshambakey, Y. Zhu, T. Hong, W. J. Tolone, S. Das Bhattacharjee, I. Cho, W. Dou, H. Wang, S. Lu, *et al.*, "Open computing infrastructure for sharing data analytics to support building energy simulations," *Journal of Computing in Civil Engineering*, vol. 33, no. 6, p. 04019037, 2019.

[130] R. Zhang and O. T. Karaguzel, "Development and calibration of reduced-order building energy models by coupling with high-order simulations," *Global journal of advanced engineering technologies and sciences*, vol. 7, no. 2, 2020.

[131] W. J Tolone, "Application of the virtual information fabric infrastructure (vifi) to building performance simulations," *Current Trends in Civil & Structural Engineering*, vol. 4, no. 2, 2019.

[132] S. D. Bhattacharjee, W. J. Tolone, A. Mahabal, M. Elshambakey, I. Cho, A. a.-R. Nayeem, J. Yuan, and G. Djorgovski, "Multi-view, generative, transfer learning for distributed time series classification," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 5585–5594, IEEE, 2019.

[133] S. Gesing, J. Krüger, R. Grunzke, S. Herres-Pawlis, and A. Hoffmann, "Using science gateways for bridging the differences between research infrastructures," *Journal of Grid Computing*, vol. 14, no. 4, pp. 545–557, 2016.

[134] I. Foster, "Globus online: Accelerating and democratizing science through cloud-based services," *IEEE Internet Computing*, vol. 15, no. 3, pp. 70–73, 2011.

[135] S. Gugnani, C. Blanco, T. Kiss, and G. Terstyanszky, "Extending science gateway frameworks to support big data applications in the cloud," *Journal of Grid Computing*, vol. 14, no. 4, pp. 589–601, 2016.

[136] I. Sfiligoi, D. C. Bradley, B. Holzman, P. Mhashilkar, S. Padhi, and F. Wurthwein, "The pilot way to grid resources using glideinwms," in *Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering - Volume 02*, CSIE '09, (Washington, DC, USA), pp. 428–432, IEEE Computer Society, 2009.

[137] A. S. Szalay, "From skyserver to sciserver," *The ANNALS of the American Academy of Political and Social Science*, vol. 675, no. 1, pp. 202–220, 2018.

[138] http://jupyter.org/.

[139] N. Wilkins-Diehr, "Special issue: Science gatewaysâcommon community interfaces to grid resources," *Concurrency and Computation: Practice and Experience*, vol. 19, no. 6, pp. 743–749, 2007.

[140] Y. Gong, L. Morandini, and R. O. Sinnott, "The design and benchmarking of a cloud-based platform for processing and visualization of traffic data," in *IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 13–20, Feb 2017.

[141] http://www.prace-ri.eu/.

[142] V. Dimitrov, "Evolution of the european grid infrastructure from grid to cloud," *Proceedings of International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE)*, p. 610, 2013. Date revised - 2014-12-01; Last updated - 2015-01-06.

[143] T. Gottdank, *Introduction to the WS-PGRADE/gUSE Science Gateway Framework*, pp. 19–32. Cham: Springer International Publishing, 2014.

[144] T. Piontek, B. Bosak, M. Ciżnicki, P. Grabowski, P. Kopta, M. Kulczewski, D. Szejnfeld, and K. Kurowski, "Development of science gateways using qcg — lessons learned from the deployment on large scale distributed and hpc infrastructures," *Journal of Grid Computing*, vol. 14, pp. 559–573, Dec 2016.

[145] Z. Farkas, P. Kacsuk, and Á. Hajnal, "Enabling workflow-oriented science gateways to access multi-cloud systems," *Journal of Grid Computing*, vol. 14, pp. 619–640, Dec 2016.

[146] S. Gesing, K. Lawrence, M. Dahan, M. E. Pierce, N. Wilkins-Diehr, and M. Zentner, "Science gateways: Sustainability via on-campus teams," *Future Generation Computer Systems*, vol. 94, pp. 97 – 102, 2019.

[147] S. Gesing, J. KrÃŒger, R. Grunzke, S. Herres-Pawlis, and A. Hoffmann, "Challenges and modifications for creating a mosgrid science gateway for us and european infrastructures," in *7th International Workshop on Science Gateways*, pp. 73–79, June 2015.

[148] J. Arshad, G. Terstyanszky, T. Kiss, N. Weingarten, and G. Taffoni, "A formal approach to support interoperability in scientific meta-workflows," *Journal of Grid Computing*, vol. 14, pp. 655–671, Dec 2016.

[149] S. Gugnani, C. Blanco, T. Kiss, and G. Terstyanszky, "Extending science gateway frameworks to support big data applications in the cloud," *Journal of Grid Computing*, vol. 14, pp. 589–601, Dec 2016.

[150] Z. Farkas, P. Kacsuk, and A. Hajnal, "Connecting workflow-oriented science gateways to multi-cloud systems," in *7th International Workshop on Science Gateways*, pp. 40–46, June 2015.

[151] M. Pierce, S. Marru, L. Gunathilake, T. A. Kanewala, R. Singh, S. Wijeratne, C. Wimalasena, C. Herath, E. Chinthaka, C. Mattmann, A. Slominski, and P. Tangchaisin, "Apache airavata: Design and directions of a science gateway framework," in *6th International Workshop on Science Gateways*, pp. 48–54, June 2014.

[152] http://developer.agaveapi.co/.

[153] I. Foster, "Globus online: Accelerating and democratizing science through cloud-based services," *IEEE Internet Computing*, vol. 15, pp. 70–73, May 2011.

[154] K. Chard, S. Tuecke, and I. Foster, "Globus: Recent enhancements and future plans," in *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale*, XSEDE16, (New York, NY, USA), pp. 27:1–27:8, ACM, 2016.

[155] B. Allen, J. Bresnahan, L. Childers, I. Foster, G. Kandaswamy, R. Kettimuthu, J. Kordas, M. Link, S. Martin, K. Pickett, and S. Tuecke, "Software as a service for data scientists," *Commun. ACM*, vol. 55, pp. 81–88, Feb. 2012.

[156] R. T. Fielding, *Architectural styles and the design of network-based software architectures*. University of California, Irvine, 2000.

[157] C. Anderson, "Docker [software engineering]," *IEEE Software*, vol. 32, pp. 102–c3, May 2015.

[158] D. Bernstein, "Containers and cloud: From lxc to docker to kubernetes," *IEEE Cloud Computing*, vol. 1, pp. 81–84, Sept 2014.

[159] C. Boettiger, "An introduction to docker for reproducible research," *SIGOPS Oper. Syst. Rev.*, vol. 49, pp. 71–79, Jan. 2015.

[160] I. Miell and A. H. Sayers, *Docker in Practice*. Greenwich, CT, USA: Manning Publications Co., 1st ed., 2016.

[161] https://nifi.apache.org/.

[162] https://docs.docker.com/engine/swarm/.

[163] A. MÄtÄcuÅ£Ä and C. Popa, "Big data analytics: Analysis of features and performance of big data ingestion tools," *Informatica Economica*, vol. 22, no. 2, pp. 25–34, 2018.

[164] P. Kacsuk, *Science gateways for distributed computing infrastructures: Development framework and exploitation by scientific user communities*. Springer International Publishing, 8 2014.

[165] N. Shakhovska, N. Boyko, Y. Zasoba, and E. Benova, "Big data processing technologies in distributed information systems," *Procedia Computer Science*, vol. 160, pp. 561–566, 2019. The 10th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2019) / The 9th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2019) / Affiliated Workshops.

[166] https://www.incommon.org/.

[167] https://wso2.com/identity-server/.

[168] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, (Sydney, Australia), 2013.

[169] https://hub.docker.com/.

[170] G. M. Kurtzer, V. Sochat, and M. W. Bauer, "Singularity: Scientific containers for mobility of compute," *PLOS ONE*, vol. 12, pp. 1–20, 05 2017.

[171] C. Arango, R. Dernat, and J. Sanabria, "Performance evaluation of container-based virtualization for high performance computing environments," *CoRR*, vol. abs/1709.10140, 2017.

[172] E. Le and D. Paz, "Performance analysis of applications using singularity container on sdsc comet," in *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*, PEARC17, (New York, NY, USA), pp. 66:1–66:4, ACM, 2017.

[173] N. Naik, "Building a virtual system of systems using docker swarm in multiple clouds," in *IEEE International Symposium on Systems Engineering (ISSE)*, pp. 1–3, Oct 2016.

[174] https://www.shbe.org/.

[175] A. M. Wootten, E. C. Massoud, A. Sengupta, D. E. Waliser, and H. Lee, "The effect of statistical downscaling on the weighting of multi-model ensembles of precipitation," *Climate*, vol. 8, no. 12, 2020.

[176] K. Jacobs, *The US national climate assessment*. New York, NY: Springer Berlin Heidelberg, 2016.

[177] H. Butler, M. Daly, A. Doyle, S. Gillies, S. Hagen, T. Schaub, *et al.*, "The geojson format," *Internet Engineering Task Force (IETF)*, 2016.

[178] M. Vartak, S. Huang, T. Siddiqui, S. Madden, and A. Parameswaran, "Towards visualization recommendation systems," *ACM SIGMOD Record*, vol. 45, no. 4, pp. 34–39, 2017.

[179] Y. Wang, "An open source software suite for multi-dimensional meteorological data computation and visualisation," *Journal of Open Research Software*, vol. 7, 07 2019.

[180] X. Chen, L. Shen, Z. Sha, R. Liu, S. Chen, G. Ji, and C. Tan, "A survey of multi-space techniques in spatio-temporal simulation data visualization," *Visual Informatics*, vol. 3, no. 3, pp. 129–139, 2019.

[181] P. B. Gibson, S. E. Perkins-Kirkpatrick, P. Uotila, A. S. Pepler, and L. V. Alexander, "On the use of self-organizing maps for studying climate extremes," *Journal of Geophysical Research: Atmospheres*, vol. 122, no. 7, pp. 3891–3903, 2017.

[182] D. Reidmiller, C. Avery, D. Easterling, K. Kunkel, K. Lewis, T. Maycock, and B. Stewart, "Fourth national climate assessment," *Volume II: Impacts, Risks, and Adaptation in the United States*, 2017.

[183] H. Lee, D. E. Waliser, R. Ferraro, T. Iguchi, C. D. Peters-Lidard, B. J. Tian, P. C. Loikith, and D. B. Wright, "Evaluating hourly rainfall characteristics over the us great plains in dynamically downscaled climate model simulations using nasa-unified wrf," *Journal of Geophysical Research-Atmospheres*, vol. 122, no. 14, pp. 7371–7384, 2017.

[184] F. De Sales and Y. K. Xue, "Dynamic downscaling of 22-year cfs winter seasonal hindcasts with the ucla-eta regional climate model over the united states," *Climate Dynamics*, vol. 41, no. 2, pp. 255–275, 2013.

[185] P. R. Keller, M. M. Keller, S. Markel, A. J. Mallinckrodt, and S. McKay, "Visual cues: practical data visualization," *Computers in Physics*, vol. 8, no. 3, pp. 297–298, 1994.

[186] E. L. Koua, A. Maceachren, and M. â. Kraak, "Evaluating the usability of visualization methods in an exploratory geovisualization environment," *International Journal of Geographical Information Science*, vol. 20, no. 4, pp. 425–448, 2006.

[187] E. Mayr, G. Schreder, S. Salisu, and F. Windhager, "Integrated visualization of space and time: A distributed cognition perspective," 2018.

[188] P. Ogao and M.-J. Kraak, "Defining visualization operations for temporal cartographic animation design," *International Journal of Applied Earth Observation and Geoinformation*, vol. 4, no. 1, pp. 23–31, 2002.

[189] C. Beale, H. Norouzi, Z. Sharifnezhadazizi, A. R. Bah, P. Yu, Y. Yu, R. Blake, A. Vaculik, and J. Gonzalez-Cruz, "Comparison of diurnal variation of land surface temperature from goes-16 abi and modis instruments," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 4, pp. 572–576, 2020.

[190] J.-G. Lee and M. Kang, "Geospatial big data: Challenges and opportunities," *Big Data Research*, vol. 2, no. 2, pp. 74–81, 2015. Visions on Big Data.

[191] X. Huang, "Supporting location-based services in spatial network databases," in *Handbook of Research on Innovations in Database Technologies and Applications: Current and Future Trends*, pp. 316–324, IGI Global, 2009.

[192] N. Pelekis, B. Theodoulidis, I. Kopanakis, and Y. Theodoridis, "Literature review of spatio-temporal database models," *The Knowledge Engineering Review*, vol. 19, no. 3, pp. 235–274, 2004.

[193] S. Shekhar, Z. Jiang, R. Y. Ali, E. Eftelioglu, X. Tang, V. M. Gunturi, and X. Zhou, "Spatiotemporal data mining: A computational perspective," *ISPRS International Journal of Geo-Information*, vol. 4, no. 4, pp. 2306–2338, 2015.

[194] G. Percivall, "The power of location." https://www.ogc.org/blog/1817, accessed October 10, 2022.

[195] M. Hugentobler, *Quantum GIS*, pp. 935–939. Boston, MA: Springer US, 2008.

[196] D. N. Williams, R. S. Drach, P. F. Dubois, C. Doutriaux, C. J. OâConnor, K. M. AchutaRao, and M. Fiorino, "Climate data analysis tool: An open software system approach," in *13th Symp. on Global Change and Climate Variations*, 2002.

[197] D. Meyer and M. Riechert, "Open source qgis toolkit for the advanced research wrf modelling system," *Environmental Modelling & Software*, vol. 112, pp. 166–178, 2019.

[198] M. A. Rico-Ramirez, I. D. Cluckie, G. Shepherd, and A. Pallot, "Meteoinfo: Gis software for meteorological data visualization and analysis," *Meteorological Applications*, vol. 14, pp. 117–129, 2007.

[199] G. M. H. Zahan, D. Mondal, and C. Gutwin, "Contour line stylization to visualize multivariate information," in *Graphics Interface 2021*, 2021.

[200] T. Nagel, E. Duval, and A. Vande Moere, "Interactive exploration of geospatial network visualization," in *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, (New York, NY, USA), p. 557â572, Association for Computing Machinery, 2012.

[201] T. Mahmood, W. Fulmer, N. Mungoli, J. Huang, and A. Lu, "Improving information sharing and collaborative analysis for remote geospatial visualization using mixed reality," in *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 236–247, IEEE, 2019.

[202] OpenStreetMap, "Openstreetmap." https://www.openstreetmap.org/, accessed March 31, 2021.

[203] T. Hahmann and E. L. Usery, "What is in a contour map?," in *International Conference on Spatial Information Theory*, pp. 375–399, Springer, 2015.

[204] S. Wehrend and C. Lewis, "A problem-oriented classification of visualization techniques," in *Proceedings of the First IEEE Conference on Visualization: Visualization '90*, pp. 139–143, 1990.

[205] J. Heer and M. Bostock, "Crowdsourcing graphical perception: Using mechanical turk to assess visualization design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, (New York, NY, USA), p. 203â212, Association for Computing Machinery, 2010.

[206] A. Karduni, R. Wesslen, I. Cho, and W. Dou, "Du bois wrapped bar chart: Visualizing categorical data with disproportionate values," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, (New York, NY, USA), p. 1â12, Association for Computing Machinery, 2020.

[207] A. Karduni, D. Markant, R. Wesslen, and W. Dou, "A bayesian cognition approach for belief updating of correlation judgement through uncertainty visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 978–988, 2021.

[208] J. Chandler, C. Rosenzweig, A. J. Moss, J. Robinson, and L. Litman, "Online panels in social science research: Expanding sampling methods beyond mechanical turk," *Behavior research methods*, vol. 51, no. 5, pp. 2022–2038, 2019.

[209] L. St, S. Wold, *et al.*, "Analysis of variance (anova)," *Chemometrics and intelligent laboratory systems*, vol. 6, no. 4, pp. 259–272, 1989.

[210] S. Morgenthaler, "Exploratory data analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, pp. 33–44, 2009.

[211] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, "Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization," *Geoscientific Model Development*, vol. 9, no. 5, pp. 1937–1958, 2016.

[212] M. Stockhause, R. Matthews, A. Pirani, A. M. Treguier, and O. Yelekci, "Cmip6 data documentation and citation in ipcc's sixth assessment report (ar6)," in *EGU General Assembly Conference Abstracts*, pp. EGU21–2886, 2021.

[213] T. Nocke, T. Sterzel, M. BÃ¶ttinger, and M. Wrobel, "Visualization of climate and climate change data: An overview."

[214] J. D. Blower, A. L. Gemmell, G. H. Griffiths, K. Haines, A. Santokhee, and X. Yang, "A web map service implementation for the visualization of multidimensional gridded environmental data," *Environmental Modelling and Software*, vol. 47, pp. 218–224, 9 2013.

[215] R. R. Nemani, W. Wang, A. Michaelis, P. Votava, and S. Ganguly, "OpenNEX, a private-public partnership in support of the national climate assessment," in *AGU Fall Meeting Abstracts*, vol. 2016, pp. GC31G–1182, Dec. 2016.

[216] A. Talukder, M. Elshambakey, S. Wadkar, H. Lee, L. Cinquini, S. Schlueter, I. Cho, W. Dou, and D. J. Crichton, "Vifi: Virtual information fabric infrastructure for data-driven discoveries from distributed earth science data; vifi: Virtual information fabric infrastructure for data-driven discoveries from distributed earth science data," 2017.

[217] G. A. Meehl, W. M. Washington, J. M. Arblaster, A. X. Hu, H. Y. Teng, J. E. Kay, A. Gettelman, D. M. Lawrence, B. M. Sanderson, and W. G. Strand, "Climate change projections in cesm1(cam5) compared to ccsm4," *Journal of Climate*, vol. 26, no. 17, pp. 6287–6308, 2013.

[218] M. Kelley, G. A. Schmidt, L. S. Nazarenko, S. E. Bauer, R. Ruedy, G. L. Russell, A. S. Ackerman, I. Aleinov, M. Bauer, R. Bleck, *et al.*, "Giss-e2. 1: Configurations and climatology," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 8, p. e2019MS002025, 2020.

[219] N. Kaye, A. Hartley, and D. Hemming, "Mapping the climate: guidance on appropriate techniques to map climate variables and their uncertainty," *Geoscientific Model Development*, vol. 5, no. 1, pp. 245–256, 2012.

[220] A. Sarikaya, M. Gleicher, and D. A. Szafir, "Design factors for summary visualization in visual analytics," in *Computer Graphics Forum*, vol. 37, pp. 145–156, Wiley Online Library, 2018.

[221] U. Rebbapragada, P. Protopapas, C. E. Brodley, and C. Alcock, "Finding anomalous periodic time series," *Machine learning*, vol. 74, no. 3, pp. 281–313, 2009.

[222] J. Meyer, E. W. Bethel, J. L. Horsman, S. S. Hubbard, H. Krishnan, A. Romosan, E. H. Keating, L. Monroe, R. Strelitz, P. Moore, *et al.*, "Visual data analysis as an integral part of environmental management," *IEEE transactions on visualization and computer graphics*, vol. 18, no. 12, pp. 2088–2094, 2012.

[223] J. Johansson, T.-S. S. Neset, and B.-O. Linnér, "Evaluating climate visualization: An information visualization approach," in *2010 14th International Conference Information Visualisation*, pp. 156–161, IEEE, 2010.

[224] N. Elmqvist and J. S. Yi, "Patterns for visualization evaluation," *Information Visualization*, vol. 14, no. 3, pp. 250–269, 2015.

[225] S. V. Raghavan, J. Hur, and S.-Y. Liong, "Evaluations of nasa nex-gddp data over southeast asia: present and future climates," *Climatic change*, vol. 148, no. 4, pp. 503–518, 2018.

[226] G. Jiang, C. Zhao, M. R. Scott, and F. Zou, "Combinable tabs: An interactive method of information comparison using a combinable tabbed document interface," in *IFIP Conference on Human-Computer Interaction*, pp. 432–435, Springer, 2009.

[227] E. Zgraggen, A. Galakatos, A. Crotty, J.-D. Fekete, and T. Kraska, "How progressive visualizations affect exploratory analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 8, pp. 1977–1987, 2017.

[228] T. Schreck, D. Keim, and F. Mansmann, "Regular treemap layouts for visual analysis of hierarchical data," in *Proceedings of the 22nd Spring Conference on Computer Graphics*, SCCG '06, (New York, NY, USA), p. 183â190, Association for Computing Machinery, 2006.

[229] K. Marriott, F. Schreiber, T. Dwyer, K. Klein, N. H. Riche, T. Itoh, W. Stuerzlinger, and B. H. Thomas, *Immersive Analytics*, vol. 11190. Springer, 2018.

[230] T. Sellis, A. D. Library., and A. for Computing Machinery. Special Interest Group on Management of Data., *An Analytic Data Engine for Visualization in Tableau*. ACM, 2011.

[231] D. Cao, J. Zhang, L. Xun, S. Yang, J. Wang, and F. Yao, "Spatiotemporal variations of global terrestrial vegetation climate potential productivity under climate change," *Science of The Total Environment*, vol. 770, p. 145320, 2021.

[232] M. E. Porter, M. C. Hill, T. Harris, A. Brookfield, and X. Li, "The discoverframework freeware toolkit for multivariate spatio-temporal environmental data visualization and evaluation," *Environmental Modelling & Software*, vol. 143, p. 105104, 2021.

[233] C. Tominski, J. F. Donges, and T. Nocke, "Information visualization in climate research," in *2011 15th International Conference on Information Visualisation*, pp. 298–305, July 2011.

[234] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, "Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization," *Geoscientific Model Development*, vol. 9, no. 5, pp. 1937–1958, 2016.

APPENDIX A: MATERIALS FROM THE DOMAIN EXPERT SURVEY OF GEOVISUALIZATION REQUIREMENTS AND CURRENT PRACTICES FROM CLIMATE SCIENTISTS

This appendix contains the materials used in a domain expert survey with IRB approval (IRB Number: 22-0001), to identify the usability and requirements for interactive geospatial visualization. The survey invited domain experts in earth science to outline their current practices and potential gaps in geovisual analysis. The items leveraged in the survey are included in the order as follows:

1. Electronic Advertisement for Recruitment
2. Informed Consent Form
3. Demographic Information & Survey Questionnaire

Surveying Domain Expert's Knowledge on Geospatial Data Analysis

We are researchers at the University of North Carolina at Charlotte seeking participants for our research study on "A Comparative Analysis of Evaluating Interactive Geospatial Visualization". To participate in this study, we were looking for experts in the earth science domain with an understanding of geospatial data analysis, aged over 18 and proficient in English language.

In this study, we will first ask the participants a set of questionnaires using a Google Form about their work with earth science data and the tools that they currently use to accomplish the tasks. The purpose of this research study is to assist earth science, researchers and analysts, in their exploration and sense-making tasks with the interactive features in geospatial visualization. In this study, we want to better understand your analytical tasks, who are experts in the earth science domain so that we can design an interactive visual component to support your analysis tasks.

After the initial set of questionnaires about their qualification and experience in analyzing geospatial data, we will ask detailed questions about the analysis tasks and how you perform. Our main goal is to understand and categorize the analysis tasks a researcher or analyst performs on a daily basis with geospatial data. Obtaining your domain knowledge would help us to design an interactive geospatial visualization that supports your analysis tasks through visual analysis. Moreover, we would like to mention that we are not capturing any interaction log during the survey. The online survey will take approximately 30 minutes to complete.

We will provide a consent form at the beginning of the survey to explain the detailed procedure of our study. If you agree to provide your consent, the form will open the survey questionnaire to you.

Thank you and we look forward to your participation!

Project Investigator:
Abdullah-Al-Raihan Nayeem
Research Assistant
Email: anayeem@uncc.edu

Faculty Advisor:
Dr. Isaac Cho
Assistant Professor
Email: Isaac.Cho@uncc.edu

College of Computing & Informatics
9201 University City Boulevard, Charlotte, NC 28223-0001

Consent to Participate in a Research Study

Title of the Project: A Comparative Analysis of Evaluating Interactive Geospatial Visualization
Principal Investigator: Abdullah-Al-Raihan Nayeem, Research Assistant, UNC Charlotte
Faculty Advisor: Dr. Isaac Cho, Adjunct Professor, UNC Charlotte
Study Sponsor: -

You are invited to participate in a research study. Participation in this research study is voluntary. The information provided is to help you decide whether or not to participate. If you have any questions, please ask.

Important Information You Need to Know

- You are being asked to participate in a survey for a research project, "Comparative Analysis of Evaluating Interactive Geospatial Visualization". The purpose of this research study is to assist earth science, researchers and analysts, in their exploration and sense-making tasks with the interactive features in geospatial visualization. In this study, we want to better understand your analytical tasks, who are experts in the earth science domain so that we can design an interactive visual component to support your analysis tasks.

- We are asking experts in the domain of earth science to participate in our study. With your consent, we will send you a link to a Google Form containing survey questions to learn about your experience working with geospatial data, the analysis task that you perform and the tools that you utilize.

- In the survey, you may choose to skip a question you do not want to answer. You may personally benefit from taking part in this research (given the visualization component we are aiming to design is useful for replacing your analysis approach) but more importantly, your input in this study may help us better understand how such visual components can assist in the earth science analysis tasks.

- Please read this form and ask any questions you may have before you decide whether to participate in this research study.

Why are we doing this study?
In this study, we intend to learn from you, the domain expert, about the analysis tasks with the geospatial data. Additionally, we want to understand the design requirements of an interactive geospatial visualization to eventually support those tasks through visual analysis..

Why are you being asked to be in this research study?
You are being asked to be in this study because you are an expert in the earth science domain and use geospatial data in the analysis tasks in your daily work.

What will happen if I take part in this study?
If you choose to participate you will complete a survey on Google form. In this study, we will first ask the participants a set of questionnaires using a Google Form about their work experience with geospatial data and the analysis tools that they currently use for the tasks. After the initial set of questionnaires about their qualification and experience in analyzing geospatial data, we will ask detailed questions about the

analysis tasks and how you perform. Our main goal is to understand and categorize the analysis tasks a researcher or analyst performs on a daily basis. Obtaining your domain knowledge would help us to design an interactive geospatial visualization that supports your analysis tasks through visual analysis. Moreover, we would like to mention that we are not capturing any interaction log during the survey. The interview will take approximately 30 minutes to complete.

What benefits might I experience?

You will not benefit directly from being in this study at this moment. Eventually, we expect our visualization tool to assist in your analysis tasks.

What risks might I experience?

We do not expect you to go through any mental or physical risk participating in this study.

How will my information be protected?

You are asked to provide your email address as part of this study. We will use your email address to send the survey Google Form after you provide your consent. The identifier email or name will be separated from the obtained survey information. The data will be accessible only to those who are working on evaluating our proposed system. All the information collected from the study will be stored in a secure server protected by the password with minimal accessibility. If google form captures the email address of the participant on its own, we will separate that information from the responses and store only the answers provided against our questionnaires and tag the response with Expert 1, 2…Expert N.

How will my information be used after the study is over?

After this study is complete, the person involved in this research study can still utilize your feedback for other studies but will not share with any other personnel. The portion of information utilized will not contain any of your identity or identifiable information. Since we are not capturing any of your interaction data in this study rather your observation and experience, we will not get back to you for further revision of this information.

Will I receive an incentive for taking part in this study?

We are not offering any incentive to participate in this study.

What other choices do I have if I don't take part in this study?

You have the total right to opt out from this study any time. If you do not want to participate in this study, please let us know. We will not send out the survey form in that case. You would not be responsible or do not require any clarification in that case.

What are my rights if I take part in this study?

It is up to you to decide to be in this research study. Participating in this study is voluntary. Even if you decide to be part of the study now, you may change your mind and stop at any time. You do not have to answer any questions you do not want to answer.

Who can answer my questions about this study and my rights as a participant?

For questions about this research, you may contact Abdullah-Al-Raihan Nayeem (anayeem@uncc.edu, 980 318 8981) or Dr. Isaac Cho (Isaac.Cho@uncc.edu). If you have questions about your rights as a research participant, or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Office of Research Protections and Integrity at 704-687-1871 or uncc-irb@uncc.edu.

Consent to Participate

By signing this document, you are agreeing to be in this study. Make sure you understand what the study is about before you sign. You will receive a copy of this document for your records. If you have any questions about the study after you sign this document, you can contact the study team using the information provided above.

I understand what the study is about and my questions so far have been answered. I agree to take part in this study.

_____

Name (PRINT)

_____

Signature Date

_____

Name & Signature of person obtaining consent Date

Domain Expert Survey on Geospatial Data Analysis

Title of the Project: A Comparative Analysis of Evaluating Interactive Geospatial Visualization
Principal Investigator: Abdullah-Al-Raihan Nayeem, Research Assistant, UNC Charlotte
Faculty Advisor: Dr. Isaac Cho, Adjunct Professor, UNC Charlotte, Assistant Professor, Utah State University.

We would like to invite you to participate in this survey for a research project, “Comparative Analysis of Evaluating Interactive Geospatial Visualization”. The purpose of this research study is to assist earth science, researchers and analysts, in their exploration and sense-making tasks with the interactive features in geospatial visualization. In this survey, we want to better understand your analytical tasks, who are experts in the earth science domain so that we can design an interactive visual component to support your analysis tasks.

For questions about this survey, you may contact Abdullah-Al-Raihan Nayeem (anayeem@uncc.edu) or Dr. Isaac Cho (Isaac.Cho@uncc.edu).

Participant's Experience

1. What is your academic background?

   Mark only one oval.

   - ◯ Undergraduate
   - ◯ Masters
   - ◯ Doctorate
   - ◯ Other

2. How do you rate your expertise with geospatial data analysis

Mark only one oval.

1
2
3
4
5

3. What position are you holding at your current job?

4. What is your major research area? (eg. climate science, geography, meteorology)

5. How many years of experience do you have working with geospatial data?

Mark only one oval.

- 1 - 3 years
- 4 - 5 years
- 5 - 10 years
- 10+ years

6. How useful geospatial visualization is in your day-to-day tasks?

7. How do you visualize the geospatial data (tools/software)?

Geospatial analysis tasks

8. What type of information do you extract from the geospatial analysis (Please provide a list)?

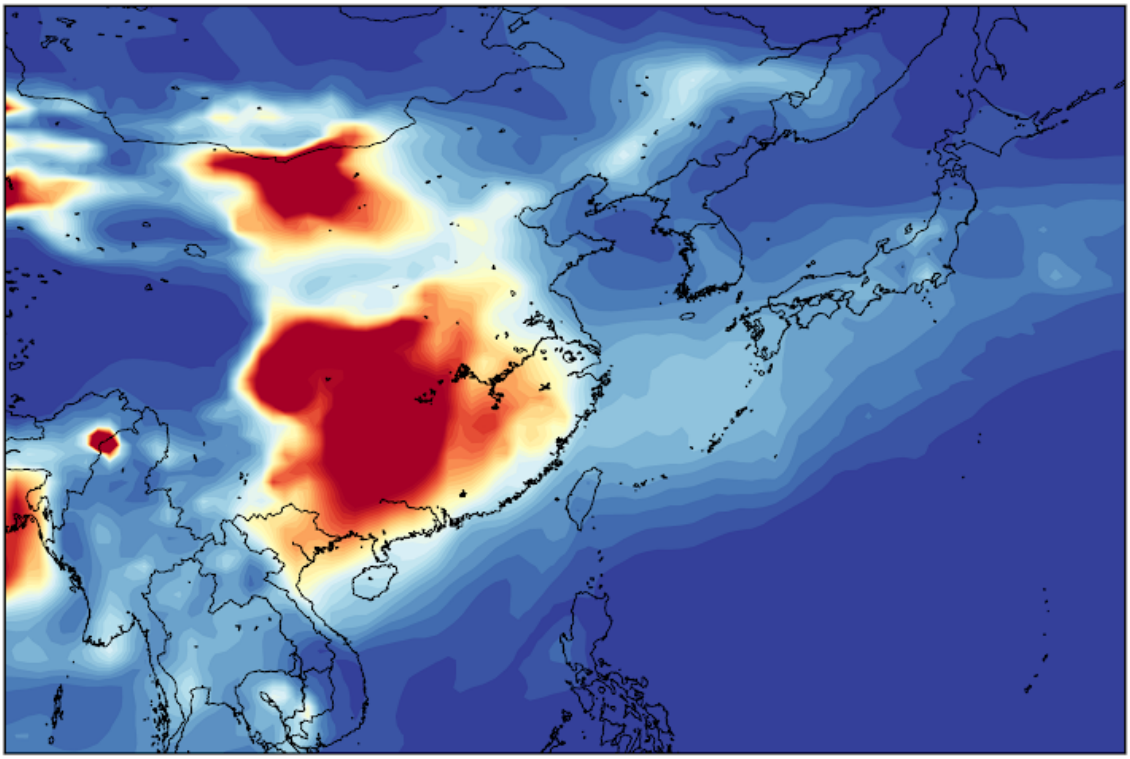
9. What type of analysis do you conduct with the geospatial data (Please provide a list)?

10. How do you compare the temporal variables in geospatial data?

11. What is the most challenging part in your geospatial analysis?

12. Which part of your analysis task would you prefer to replace with a tool?

13. How does geospatial contour map help in your analysis tasks?



APPENDIX B: MATERIALS FROM THE USER STUDY FOR QUANTITATIVE EVALUATION OF CONTOUR-BASED INTERACTIVE GEOSPATIAL VISUALIZATION FOR EXPLORATORY ANALYSIS

This appendix contains the materials used in a user intervention study to quantitatively evaluate contour-based interactive geospatial visualization, discussed in Chapter 4. This crowd-sourced study was conducted with IRB approval (IRB Number: 22-470) and recruited participants from Amazon Mechanical Turk (MTurk) for Human Intelligent Tasks (HIT). The items leveraged in the study are included in the order below:

1. Informed Consent Form
2. Posting for Amazon Mturk HIT
3. Participant Training Materials

College of Computing & Informatics
9201 University City Boulevard, Charlotte, NC 28223-0001

Consent to Participate in a Research Study

Title of the Project: A user study for evaluating interactive features on geospatial visualizations
Principal Investigator: Abdullah-Al-Raihan Nayeem, Research Assistant, UNC Charlotte
Faculty Advisor: Dr. Isaac Cho, Adjunct Professor, UNC Charlotte
Study Sponsor: -

You are invited to participate in a research study. Participation in this research study is voluntary. The information provided is to help you decide whether or not to participate. If you have any questions, please ask.

Important Information You Need to Know

- You are being asked to participate in a user study for a research project, "Quantitative Evaluation for interactive features on geospatial visualizations". The purpose of this research study is to assist earth science, researchers and analysts, in their exploration and sense-making tasks with the interactive features in geospatial visualization. In this study, we aim to conduct a user intervention study to capture users' interaction in performing their analysis tasks using geospatial visualization. We want to analyze the data to evaluate the interactive features in different analysis tasks with geospatial visualization.

- We are asking experts and students who are interested in interactive geospatial analysis to participate in our study. With your consent, we will redirect you to the user study interface. The interface will inquire a few demographic questions and provide an interactive exercise to be familiar with the visualization. Next, the interface will take you to the study page. The interface will prompt questions that you can answer using the interactive geospatial visualization. Finally, the interface will ask about your impression of using the visualization.

- In the demographic and post-questionnaire, you may choose to skip a question you do not want to answer. You may personally benefit from taking part in this research. We expect the visualization component is useful for performing analysis tasks but more importantly, your input in this study may help us better understand how such visual components can assist in the spatiotemporal analysis tasks.

- Please read this form and ask any questions you may have before you decide whether to participate in this research study.

Why are we doing this study?
In this study, we intend to learn if an interactive contour-based geospatial visualization may assist the user in performing exploratory analytical tasks over the geospatial data. W designed analytical tasks to perform using both traditional static and interactive geospatial visualization to evaluate our hypotheses. In addition, we want to understand your impression and experience with the implemented interactive features for the visualization.

Why are you being asked to be in this research study?
You are being asked to participate in this study because you are an expert or student related to earth science or geography domain and use geospatial data in the analysis tasks in your daily work. You are also able to participate in this study using a laptop or desktop. If you are an MTurk worker, you must be a master worker and have at least a 95% HIT approval rate.

What will happen if I take part in this study?
If you choose to participate and provide your consent on the web interface electronically, you will be redirected to the user study. We are not capturing any identifiable information in our study. The rest of the study will follow up and store the data against an unidentifiable unique ID.

The study starts with getting basic information from you such as academic qualification, gender, and expertise in geospatial analysis. Next, the interface forwards to another window that interactively introduces the developed geospatial visualization. This interactive exercise describes the features and how to utilize them. To verify your understanding, we would provide some sample questions that are not part of the study. If you are able to correctly answer the questions, the interface redirects to the actual user study. In the training phase, when the user chooses the wrong answer, the interface would assist in selecting the correct one.

The user study interface consists of three parts - a geospatial visualization panel, an information panel that describes the data put on the map, and a question/answer panel. The interface will provide 8 sets of questions to the user from 8 different task categories. Each set contains 14-15 questions. To answer these questions, the user needs to interact with the map. Patterns of questions would vary based on the task category. That entails that you might be prompted with a similar question set for different geospatial visualization. Finally, the interface prompts some usability questions to understand the participant's impression of using the visualization. The user study will take approximately 60 minutes to complete.

What benefits might I experience?
You will not benefit directly from being in this study at this moment. Eventually, we expect our visualization tool to assist in your analysis tasks in the future.

What risks might I experience?
We do not expect you to go through any mental or physical risk participating in this study.

How will my information be protected?
We do not intend to capture any of your identifiable information. The data captured in the user study are stored against a unique ID generated once you provide your consent to participate in the study. After completing the user study, data will be accessible only to those who are working on evaluating our proposed system. All the information collected from the study will be stored in a secure server protected by a password with minimal accessibility. We will not publish or disclose any granular portion of the data that might lead to the potential identification of a participant.

How will my information be used after the study is over?
After this study is complete, the person involved in this research study can still utilize your feedback for other studies but will not share it with any other personnel. The portion of information utilized will not contain any of your identity or identifiable information. We will not get back to you for further revision of this information.

Will I receive an incentive for taking part in this study?
If you are a participant from UNCC SONA, completion of this study would benefit you with 1 credit. If you are a participant from M-Turk, please use your token after completing the study to redeem an incentive of $4 in the M-Turk portal. If you are a participant from SONA, please contact the PI with your token, to earn the credit. Your tokens are not identifiable or associated with your participation in this

study. Please note that to earn the incentive you must complete the study and get a completion code from the user study interface. Additionally, we also place eight engagement questions to ensure your attentiveness during the study. If you are unable to answer all the engagement questions correctly or complete the user study portion in less than 10 minutes, we reserve the right to not award the incentive. That being said, we do not want you to get stressed about finding the correct answer. We would definitely appreciate your best attempt.

What other choices do I have if I don't take part in this study?

You have the total right to opt out of this study at any time. If you do not want to participate in this study, please let us know. You would not be responsible or do not require any clarification in that case.

What are my rights if I take part in this study?

It is up to you to decide to participate in this research study. Participating in this study is voluntary. Even if you choose to be part of the study now, you may change your mind and stop at any time. You do not have to answer any questions you do not want to answer.

Who can answer my questions about this study and my rights as a participant?

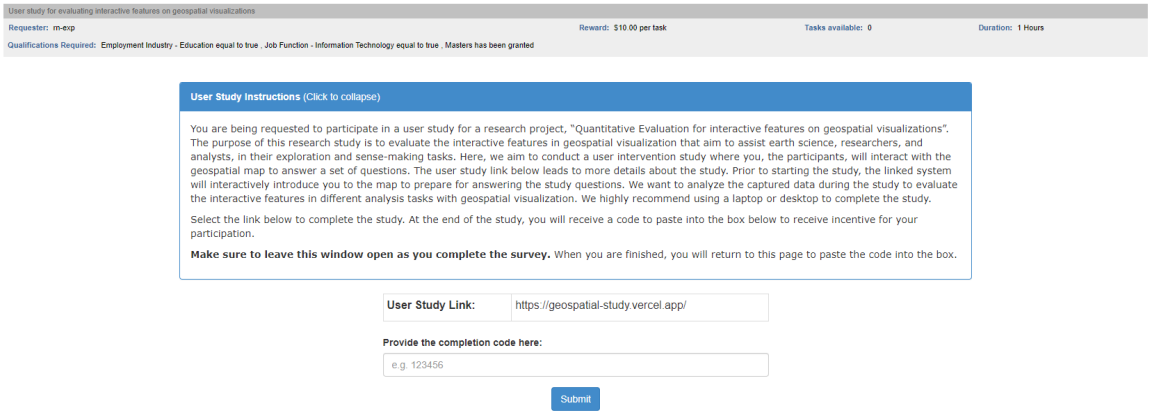
For questions about this research, you may contact Abdullah-Al-Raihan Nayeem (anayeem@uncc.edu, 980 318 8981) or Dr. Isaac Cho (Isaac.Cho@uncc.edu). If you have questions about your rights as a research participant or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Office of Research Protections and Integrity at 704-687-1871 or uncc-irb@uncc.edu.

Consent to Participate

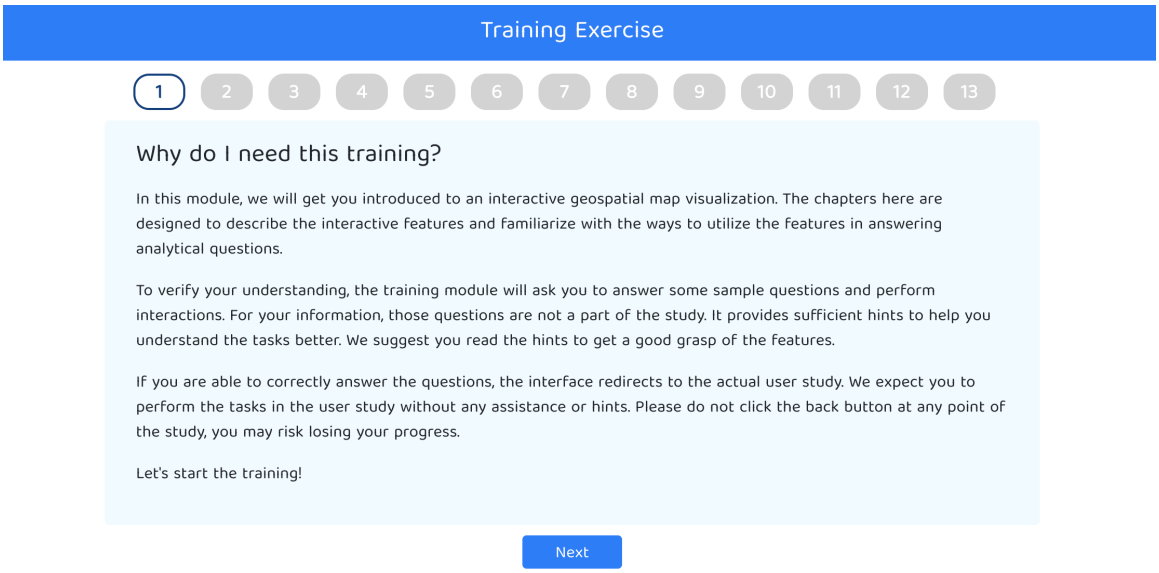
By checking the box below, you are agreeing to be in this study. Make sure you understand what the study is about before you agree. You can download a copy of this document for your records. If you have any questions about the study after you agree to provide your consent, you can contact the study team using the information provided above.

❑ I understand what the study is about and my questions so far have been answered. I agree to take part in this study.

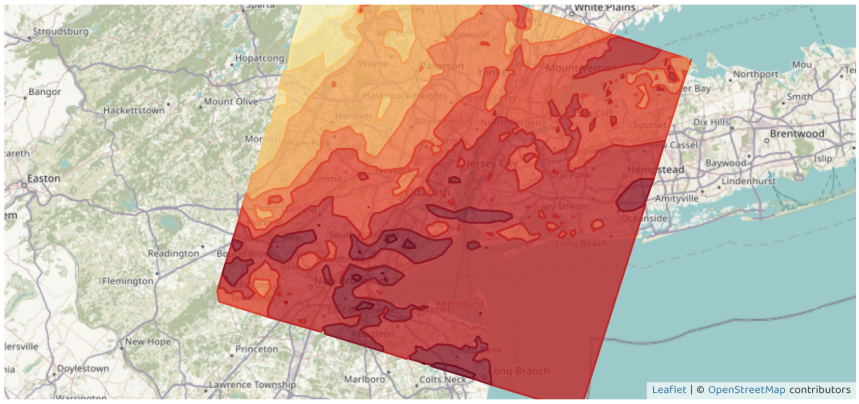
B.1 Posting for Amazon Mturk HIT



B.2 Participant Training Materials



1 2 3 4 5 6 7 8 9 10 11 12 13



Introducing Contour Map

What is contour?

Contours are lines and regions that connect places of similar values in a dataset. That indicates continuous events or measures like height, temperature, precipitation, or air pressure. The colored-boundaries within the map are identified as contour regions.

What is a contour map?

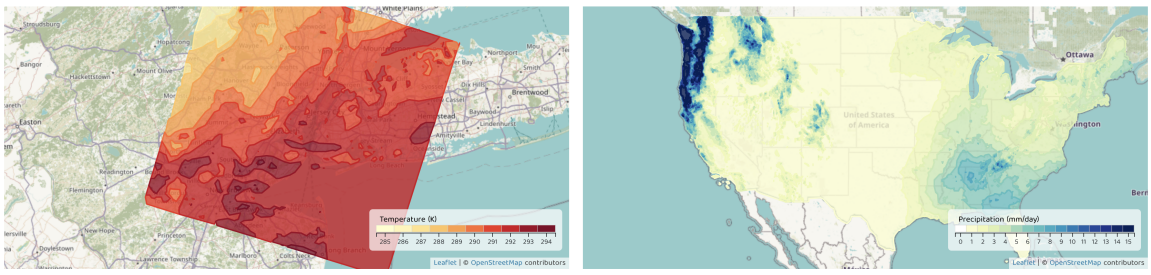
A contour map, often known as a topographic map, is a form of a geospatial map with large-scale detail and quantitative depiction of terrain characteristics (geographical features), commonly with contour lines drawn on top.

What am I looking at?

The geospatial visualization above is a **contour map of hourly-averaged temperature intensity** in New York City. Here, temperature contours are placed on the map to provide a geographical context to the view.

1 2 3 4 5 6 7 8 9 10 11 12 13

Snapshot Left | Snapshot Right



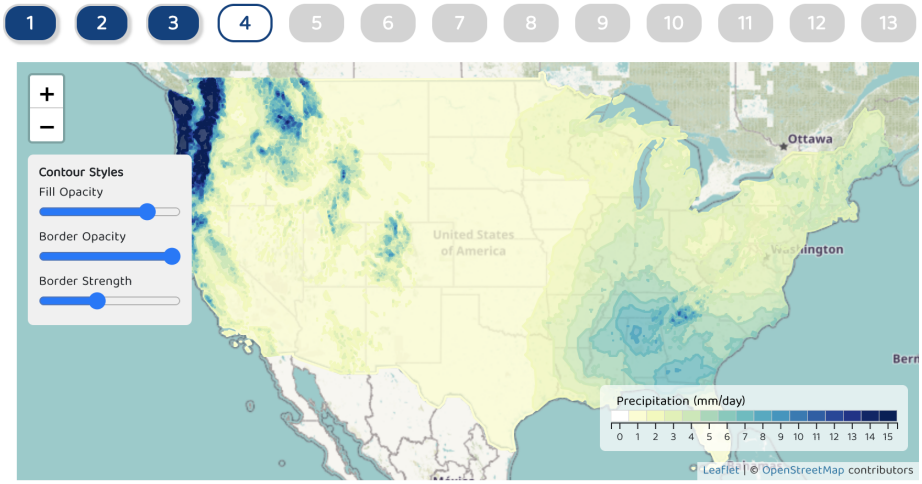
Contour Labels

Snapshot Left:
A contour map of hourly-averaged temperature intensity in the New York City.

Snapshot Right:
A contour map of monthly-averaged precipitation intensity in the Contiguous United States.

Labels:
Presents the mapping between the intensity and color. Contour map obtain different level of intensities (e.g., Precipitation ranges from 0-15 mm/day) that are labeled at the bottom right of each view. Each label (e.g, 285K, 286K, etc.) represents a unique contour region. Often cases there are multiple contour regions of an intensity on the map.

Next



Interaction Mode:: Inspect

Inspect

Moving your mouse cursor over the contour map highlights the associated contour regions on the map. On a tooltip, it provides additional information about the interacted contour region. The information includes:

[Intensity] is the temperature/precipitation recorded on the area that you are interacting with.

[Contour Region] shows two values: No. of regions with similar intensity / Total region count

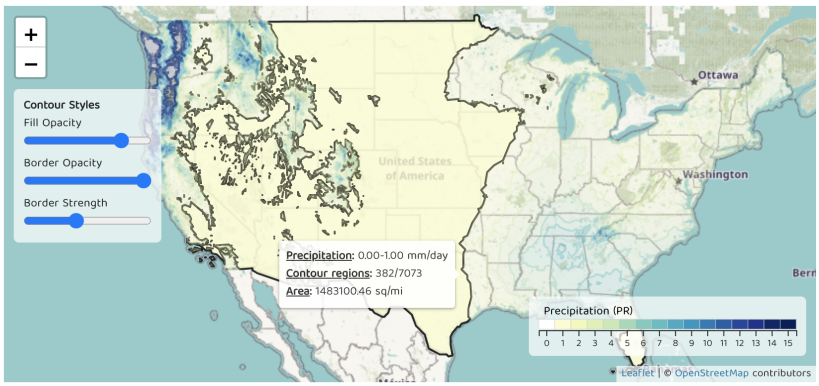
[Area] provides total area for the contour regions that are similar to the interacted region.

Contour Styles

Allows to control the transparency or strength of the contour to be able to better co-relate an event with its geographical location.

Training Exercise

1 2 3 4 5 6 7 8 9 10 11 12 13



Interaction Modes :: Inspect

- Move the mouse cursor on the geospatial contour to inspect and identify the area where the highest and lowest precipitation is recorded.

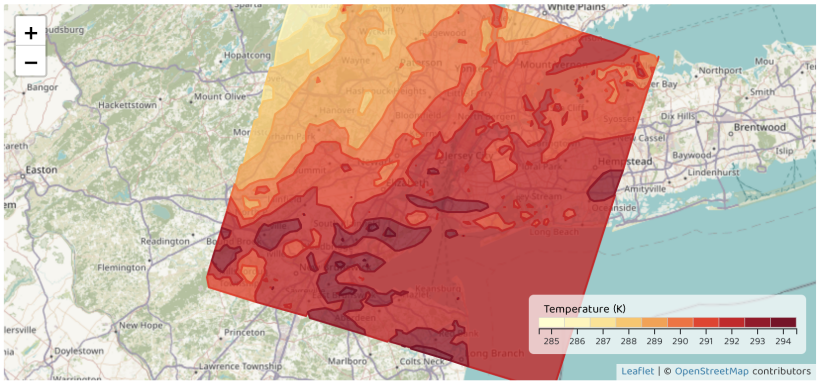
1. What is the highest precipitation on the contour?

>14.00 mm/day

2. What is the lowest precipitation on the contour?

0.00-1.00 mm/day

1 2 3 4 5 6 7 8 9 10 11 12 13



Contour Area Distribution :: Exercise

Please use the above map visualization to answer the following questions:

1. Which temperature intensity covers the largest area on the contour?

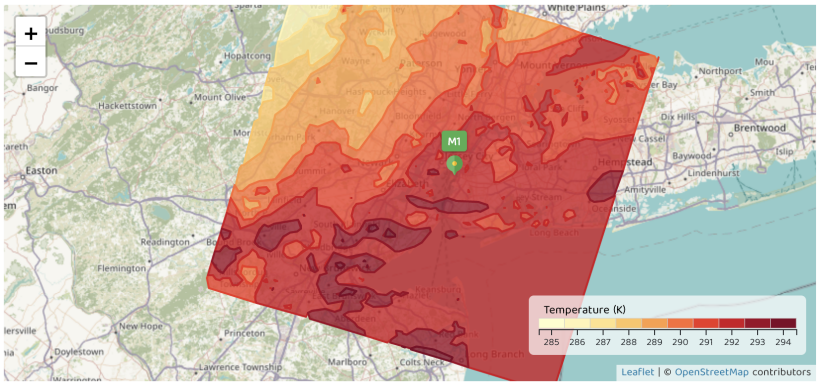
Select an option

2. Which temperature intensity covers the least area on the contour?

Select an option

Training Exercise

1 2 3 4 5 6 7 8 9 10 11 12 13



Markers on Interactive Contour

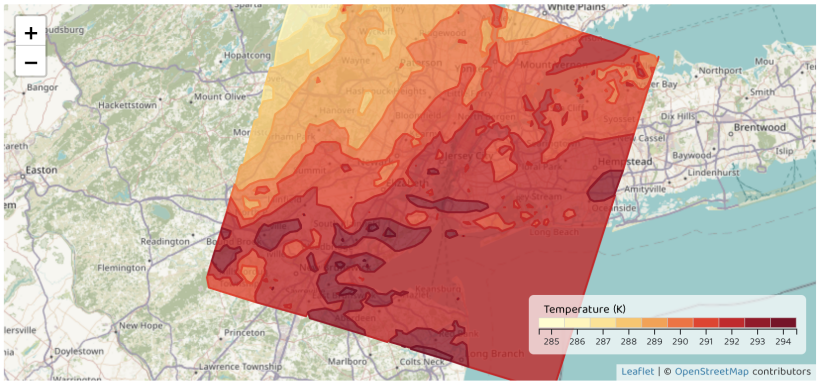
What is a marker on the contour?

A marker on the map identifies a geographic location. On the above view, a marker provides (temperature or precipitation) intensity and comparison among other markers or contour lines.

How to place a marker?

A right click on the point of interest puts a marker on the map. The view allows multiple markers in different location. A left click on the mouse shows additional information of the marked location along with the comparison with the other marked places.

1 2 3 4 5 6 7 8 9 10 11 12 13



Apply Markers on Interactive Contour

- Put a marker on the map (with a left mouse click) within the contour boundary.
- Put another two markers on different than previous intensity on the map.
Observe **M1**'s difference with the highest temperature recorded (**inspect highest temperature**).
When calculating the difference, please always consider the upper bound of the range.

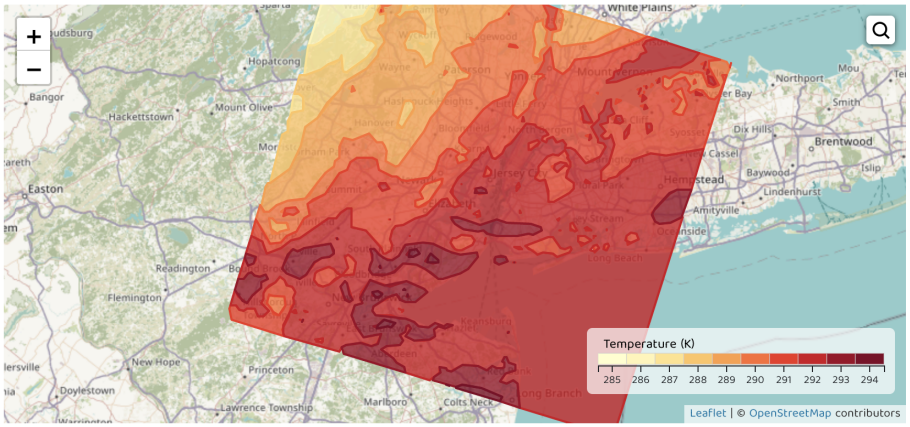
1. What is the intensity recorded in **M1**?

Select an option

2. What is the intensity recorded in **M2**?

Select an option

1 2 3 4 5 6 7 8 9 10 11 12 13

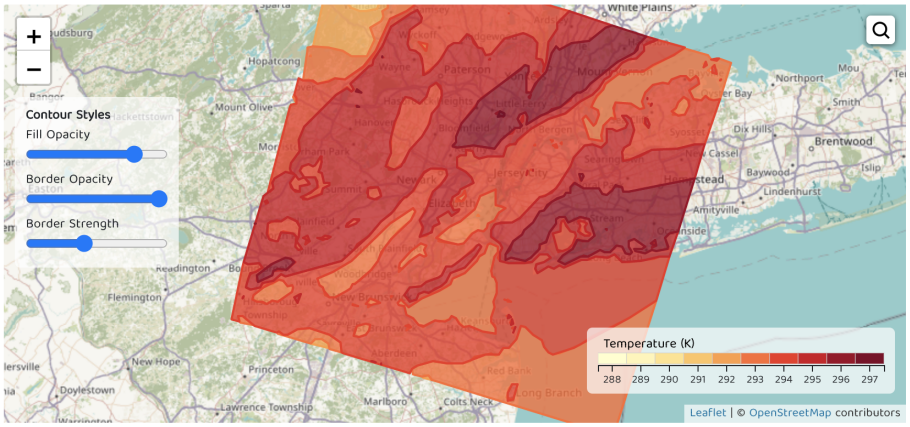


Search by Area Name within the Contour Map

- Helps to locate an area on the map that might be difficult to find manually. Consider using the search feature when you are asked about a location that you find difficult to locate.
- To search an area, click the search icon placed at the top right on the contour map. Type the area name and choose the one from the options. A marker will be placed on the searched and selected location.

**The searched markers function just as the other markers manually placed on the map.

1 2 3 4 5 6 7 8 9 10 11 12 13

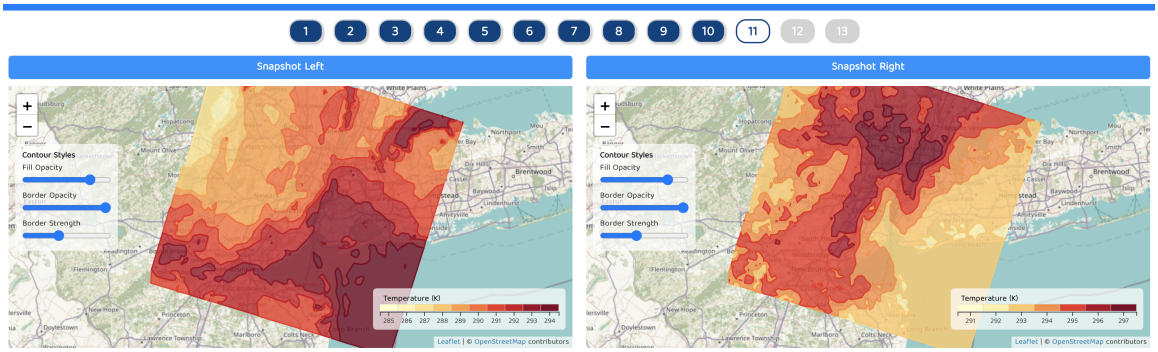


Interaction Modes :: Opacity Control

- Reduce the opacity using the 'Primary Opacity Control' in 'Interaction Mode' to make the area names visible.
- Look for 'Yonkers' and 'Newark' and put two markers on those area.

1. What is the intensity difference between two marked places?

Select an option

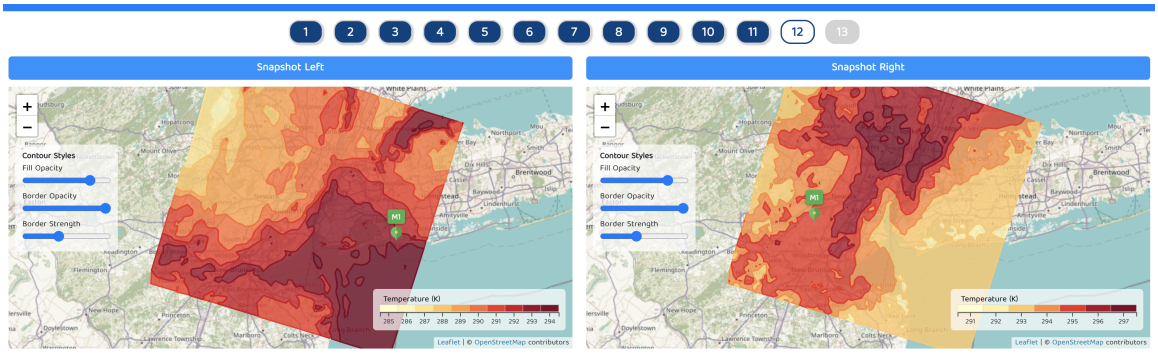


Contour Comparison

Contour Comparison
Provides exploratory features to identify the potential differences for point of interests in two different snapshots (e.g., Snapshot Left vs. Snapshot Right).

Synchronized Interactions
Interactive features are synchronized between two contour maps. That entails interaction with a map has a complementary effect on the other map.

How does marker function in this setup?
Adding markers on the map are not synchronized. However, markers can be essential to compare contour lines of different maps.



Contour Comparison :: Exercise

Markers on the right map are referenced as 'CMP Mx' on left map interactions and vice-versa.

1. What is the difference in maximum temperature between two snapshots?

Select an option

2. What is the temperature difference between M1 of these snapshots?

Select an option

1 2 3 4 5 6 7 8 9 10 11 12 13

What's next?

Awesome job! You are now well enough trained to start your study.

How does the user study differ from training?
The interface is not going to assist you with any details or hints to find the correct answer. We highly recommend you leverage the components and interact with the map to answer with your best estimation.

Are the questions going to be similar?
Yes, the questions in the study would be similar to training. However, the interface will present a different set of contour data.

How many questions do I need to answer?
There will be 8 sets each containing 12-14 questions. Please interact with the map and find your best estimation while answering a question.

Can I revise the training modules during the study?
Any time! If you need to recall the training modules, please click on the yellow circled ⚑ icon in the user study page. You can browse the modules by clicking the module numbers above.

For your Information:

- Each set of questions would be based on one/two map views.
- For a couple of sets, we expect you to answer the questions without the interactive features. The map view will be static at that time.
- We will prompt the variations of the two contours that you completed your training upon.
- Again, when calculating the differences, please always consider the upper bound of the range.
- Do not get stressed about finding the correct answer. We would appreciate your best attempt.
- Please find your closest estimation from the options while answering a question.

Let's start!

APPENDIX C: MATERIALS FROM THE USER STUDY OF VISUAL ANALYTICS APPROACHES FOR DOWNSCALED CLIMATE MODEL EXPLORATION

This appendix contains the materials used in a qualitative evaluation study with IRB approval (IRB Number: 21-0324), which is discussed in Chapter 5. The study invited domain experts in earth and climate science to evaluate our proposed visual analytics approach for large-scale spatiotemporal data using downscaled climate projections. The items below are included in the order:

1. Informed Consent Form
2. Demographic & Pre-Questionnaire Form
3. Post Questionnaire Form

College of Computing & Informatics
9201 University City Boulevard, Charlotte, NC 28223-0001

Consent to Participate in a Research Study

Title of the Project: Visual Analytics system for NASA Earth Exchange Downscaled Climate Projection
Principal Investigator: Abdullah-Al-Raihan Nayeem, Research Assistant, UNC Charlotte
Faculty Advisor: Dr. Isaac Cho, Adjunct Professor, UNC Charlotte
Study Sponsor: National Science Foundation (NSF)

You are invited to participate in a research study. Participation in this research study is voluntary. The information provided is to help you decide whether or not to participate. If you have any questions, please ask.

Important Information You Need to Know

- You are being asked to participate in a research study, "Visual Analytics System for NASA NEX-DCP30" The purpose of this research study is to assist earth science, researchers and analysts, in their exploration and sense-making tasks using the NEX-DCP30 dataset. In this study, we want to better understand your analytical tasks who are experts in the earth science domain, and how our proposed visual analytics system can contribute to your workflow.

- We are asking experts in the domain of earth science to participate in our study. This is a three (3) part study. With your consent, we will send you a link to a Google Form containing pre-questionnaire, system demonstration that includes embedded video demonstration introducing our system and opportunity to use the system without giving us any use or interaction log, and finally, post demonstration questionnaire to understand your feedback or impression about our system.

- In the pre-questionnaire section, we will ask about your work experience with earth science data and the analysis tools that you currently use for the tasks. You may choose to skip a question you do not want to answer. You may personally benefit from taking part in this research (given our system is useful for replacing your analysis tool) but more importantly, your input in this study may help us better understand how this system can assist in the earth science analysis tasks.

- Please read this form and ask any questions you may have before you decide whether to participate in this research study.

Why are we doing this study?
In this study, we intend to understand the usefulness of our system and improve in terms of features and performance to make it useful for the earth science analytical tasks.

Why are you being asked to be in this research study?
You are being asked to be in this study because you are an expert in the earth science domain and use data analysis tools in your daily work.

What will happen if I take part in this study?
If you choose to participate you will complete a survey on Google form. In this study, we will first ask the participants a set of questionnaires using a Google Form about their work experience with earth science data and the analysis tools that they currently use for the tasks. After the initial set of questionnaires, we

will provide a recorded video demonstration of our visual analytics system for earth science to facilitate analyzing the NEX-DCP30 data. Next, we will ask questions about our system based on the demonstration provided, specifically to understand the potential of the system in your future endeavors, your recommendations about the system, and usefulness of the system's features in general. In this step, we would also like the participants to use our system. We would like to mention that we are not capturing any interaction log. We do not intend to evaluate the system based on the interaction, therefore, we are not interested to know how participants use the interface. The follow-up questionnaires will also be included in the Google Form right after the video demonstration. In the process, we will ask about your visualization and interaction preferences to explore the dataset. Additionally, we will also ask your feedback on the relevancy between the analyzed data and visualization provided for exploratory analysis. The interview will take approximately 60 minutes to complete.

What benefits might I experience?
You will not benefit directly from being in this study at this moment. Eventually, we expect our system to assist in your analysis tasks.

What risks might I experience?
We do not expect you to go through any mental or physical risk participating in this study.

How will my information be protected?
You are asked to provide your email address as part of this study. We will use your email address to send the survey Google Form after you provide your consent. The identifier email or name will be separated from the obtained survey information. The data will be accessible only to those who are working on evaluating our proposed system. All the information collected from the study will be stored in a secure server protected by the password with minimal accessibility. If google form captures the email address of the participant on its own, we will separate that information from the responses and store only the answers provided against our questionnaires and tag the response with Expert 1, 2…Expert N.

How will my information be used after the study is over?
After this study is complete, the person involved in this research study can still utilize your feedback for other studies but will not share with any other personnel. The portion of information utilized will not contain any of your identity or identifiable information. Since we are not capturing any of your interaction data in this study rather your observation and experience, we will not get back to you for further revision of this information.

Will I receive an incentive for taking part in this study?
We are not offering any incentive to participate in this study.

What other choices do I have if I don't take part in this study?
You have the total right to opt out from this study any time. If you do not want to participate in this study, please let us know. We will not send out the survey form in that case. You would not be responsible or do not require any clarification in that case.

What are my rights if I take part in this study?
It is up to you to decide to be in this research study. Participating in this study is voluntary. Even if you decide to be part of the study now, you may change your mind and stop at any time. You do not have to answer any questions you do not want to answer.

Who can answer my questions about this study and my rights as a participant?
For questions about this research, you may contact Abdullah-Al-Raihan Nayeem (anayeem@uncc.edu, 980 318 8981) or Dr. Isaac Cho (Isaac.Cho@uncc.edu). If you have questions about your rights as a research participant, or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Office of Research Protections and Integrity at 704-687-1871 or uncc-irb@uncc.edu.

Consent to Participate

By signing this document, you are agreeing to be in this study. Make sure you understand what the study is about before you sign. You will receive a copy of this document for your records. If you have any questions about the study after you sign this document, you can contact the study team using the information provided above.

I understand what the study is about and my questions so far have been answered. I agree to take part in this study.

_____

Name (PRINT)

_____

Signature Date

_____

Name & Signature of person obtaining consent Date

User Experience Study :: Visual Analytics System for NASA Earth Exchange Downscaled Climate Projections

In this study, we aim to understand your analytical tasks in the earth science domain and how our proposed visual analytics system can contribute in your workflow. We would also like to receive feedback from you through the questionnaires about the currently implemented features and potential future works of our system

* Required

1. In the invitation, we have sent you a DocuSign consent form for participating in this research study. Please sign the form before proceeding to the study. *

   Mark only one oval.

   - ◯ I have signed the consent to participate in this study
   - ◯ I do not intend to participate in this study

Pre-Questionnaire

2. What is your education level?

3. What is your job position?

4. How many years of experience you have in this job?

5. Please briefly describe your work.

6. For what purpose do you study the Climate Science?

7. Which software and toolkit do you use to analyze your datasets?

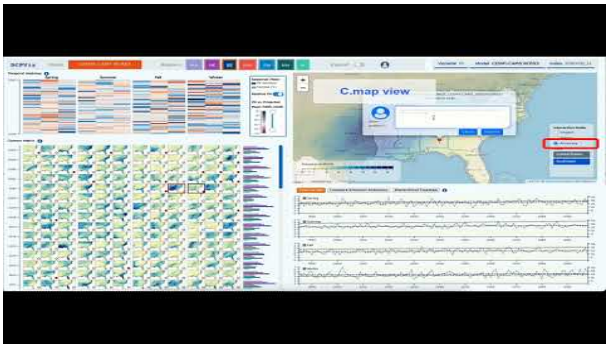
8. Which software or tool do you use to visualize your datasets?

9. What are some of the prominent datasets you generally utilize in your research and analyses?

10. How easy or difficult you find the process of data analysis and visualization?

11. Are there any specific features you wish to have when you are using visualization software (e.g. GIS + map)?

Video Demonstration	To use our interface, please visit the following link: https://esva.jpllab.net/



http://youtube.com/watch?v=INDfAZo-ZZY

Feedback on User Interface

12. Which visualization was most useful for you and why?

13. What scenario do you think you can utilize this interface to analyze your data?

14. Which features of the interface would be useful for your analysis and why?

15. Are there any specific features (or visualizations) you think to add into the interface?

16. How convenient the interface was for you to find the target information?

17. Are the visualizations self explanatory with colors, legend, and tooltip?

18. How do you rate the performance of the visualization (based on rendering time and transition)?