Stat
The ISI's Journal for the Rapid
Dissemination of Statistics Research
(wileyonlinelibrary.com) DOI: 10.100X/sta.0000

Visuanimation in statistics[+]

Marc G. Genton[a*], Stefano Castruccio[a], Paola Crippa[a], Subhajit Dutta[a], Raphaël Huser[a], Ying Sun[a], Sabrina Vettori[a]

Received February 11, 2015;

This paper explores the use of visualization through animations, coined visuanimation, in the field of statistics. In particular, it illustrates the embedding of animations in the paper itself and the storage of larger movies in the online supplemental material. We present results from statistics research projects using a variety of visuanimations, ranging from exploratory data analysis of image data sets to spatio-temporal extreme event modeling; these include a multiscale analysis of classification methods, the study of the effects of a simulated explosive volcanic eruption, and an emulation of climate model output. This paper serves as an illustration of visuanimation for future publications in Stat. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: animation; classification; extremes; functional data; spatio-temporal data; visualization

1. Introduction

In our world of increasingly complex and big data structures and models, visualization of information and of analyzed results is becoming more important than ever. If one believes the adage from the beginning of the 20th century that "a picture is worth a thousand words," then it seems appropriate at the beginning of the 21st century to claim that "a movie is worth a million words," or at least a thousand images. In this paper, we explore the use of visualization through animations, coined visuanimation, in the field of statistics. This is particularly relevant for an online journal such as Stat, where animations can be embedded in the paper itself and larger movies can be stored in the online supplemental material of the paper. Although this attractive possibility has not yet been utilized so far in papers published in the journal Stat, we envision that this practice will become the norm in future Stat papers.

This paper presents results from statistics research projects performed at King Abdullah University of Science and Technology (KAUST) in Saudi Arabia using a variety of visuanimation techniques. Exploratory data analysis is the ideal platform for generating animations: Section 2 considers the surface boxplot to visualize data sets of images and provides a visuanimation of this interactive exploratory tool. Section 3 describes a visuanimation of multiscale analysis of classification using data depth. Spatio-temporal data is an interesting setting for visuanimation: Section 4 aims at visualizing the output from simulation models and considers the formation and dispersion of particles in the stratosphere from an explosive volcanic eruption of Mt. Pinatubo in 1991. Results are presented visually via a 3D

[+]**Refer to the section 'Supporting Information' for computer and software requirements to play the movies**

[a]**CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia**
[*]**Email: marc.genton@kaust.edu.sa**

Prepared using staauth.cls [Version: 2012/05/12 v1.00]

virtual reality movie. Section 5 compares 3D global data from both a climate model and from a statistical emulator by exploring a visuanimation of the results. Finally, Section 6 visuanimates the spatio-temporal behavior of a model for extreme events, specifically extreme hourly precipitation in Switzerland. The paper ends with a brief discussion.

2. Exploratory data analysis of images with surface boxplots

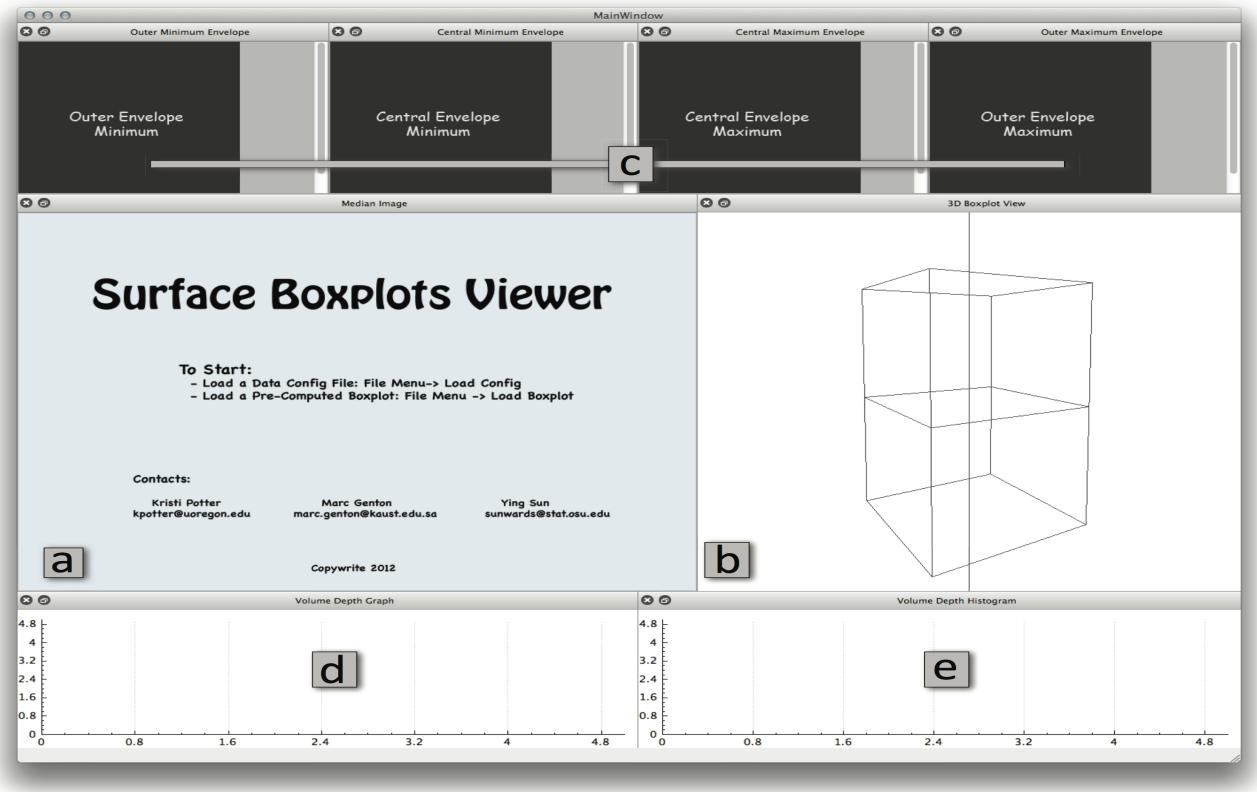
The classical boxplot is widely used as a tool to display important summary statistics for univariate variables, where observations are ordered from the smallest value to the largest in order to define the median, quartiles, the minimum and maximum, and to detect outliers. To visualize high-dimensional data or functional data of infinite dimension, simple extensions of univariate statistics to the multivariate setting do not properly capture multivariate data features because of the absence of a natural order for multi-dimensional Euclidean space. Instead, data depth provides a powerful tool for understanding multivariate ranking and quantiles. The general idea is that observations are ordered according to decreasing values of data depth (Liu et al., 1999; Zuo & Serfling, 2000). The first order statistic is defined as the median, which is the most central or the most representative observation. Subsequent order statistics induced by data depth start from the most central observation and move outwards, where larger ranks are associated with more outlying observations.

Although many notions of data depth can be used to generalize the boxplot to multivariate or functional settings, visualization remains a challenge, especially in 3D. The bagplot (Rousseeuw et al., 1999) and the functional boxplot (Sun & Genton, 2011) along with its adjusted version (Sun & Genton, 2012) are good examples of 2D visualization. The bagplot was constructed for bivariate observations using the Tukey half-space depth (Tukey, 1975), where the median data point is identified as the one with the largest depth value, and the bag is defined by the convex hull of the first 50% data points, displayed as an analog to the middle box in the classical boxplot. By contrast, the functional boxplot uses modified data depth (López-Pintado & Romo, 2009) for functional data to order sample curves, and the 50% central region is naturally defined as the envelope formed by the first 50% curves. Fast ranking by band depth is obtained by means of the algorithm proposed by Sun et al. (2012). Both the bagplot and the functional boxplot effectively summarize data features by displaying important statistics and potential outliers in 2D. Although a surface boxplot is constructed similarly, by ranking sample images, the 3D visualization can be confusing. To better display important statistics in 3D, Genton et al. (2014) developed a user-friendly visualization tool and applied it to real data examples to illustrate their features.

Movie 1 complements the description of the surface boxplot approach in Genton et al. (2014). It shows an interactive visualization tool for exploring data sets of images using the surface boxplot to extract descriptive statistics. It starts with an overview of the surface boxplot viewer, which uses a multi-window approach to display the median, inner and outer envelopes, and potential outliers. Next, we see the depth value plot and the corresponding histogram, followed by a brain fMRI data example. In addition to the multi-window display of the informative brain images, the movie also illustrates the features of image selection and identification between the depth graph and the histogram. In the movie, the corresponding image is highlighted when a point on the depth graph is selected. Similarly, if the user selects a bin on the histogram, then it is highlighted along with all corresponding images in the depth graph. The movie finishes by comparing climate model outputs from two different periods: historical run and future projection. This demonstrates through visuanimation that the surface boxplot tool is highly interactive for exploration and comparison.

3. Multiscale analysis of classification using data depth

The aim of a classification problem is to construct a rule to classify a new unlabeled observation to one of J predefined classes. Let π_i and $f_i(\mathbf{x})$ be the prior probability and the probability density function of the i-th class, respectively.



Movie 1. Interactive visualization tool for exploring data sets of images using the surface boxplot to extract descriptive statistics.

The Bayes classifier assigns an observation to the class having the maximum posterior probability. An observation $\mathbf{x}$ is classified to class i^* if

$$i^* = \arg\max_{1\le i\le J} p(i|\mathbf{x}) = \arg\max_{1\le i\le J} \pi_i f_i(\mathbf{x}),$$

where $p(i|\mathbf{x})$ is the posterior probability of the i-th $(1 \le i \le J)$ class. It is well known that the Bayes classifier is optimal because it has the lowest overall misclassification probability. However, these densities, $f_i(\mathbf{x})$s, are unknown in practice and are to be estimated from the training sample, $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}$, of the i-th class $(1 \le i \le J)$. Popular parametric approaches include linear and quadratic discriminant analyses, while kernel discriminant analysis and k-nearest neighbors are widely used nonparametric classifiers.

Data depth has emerged as a powerful methodology for data analysis with application in many areas including supervised classification. Assume that $f_1, \ldots, f_J$ are the density functions of J elliptically symmetric distributions, F_i, on $\mathbb{R}^d$; that is, $f_i(\mathbf{x}) = |\Sigma_i|^{-1/2} g_i\{\|\Sigma_i^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_i)\|\}$ for $1 \le i \le J$. Here $\boldsymbol{\mu}_i \in \mathbb{R}^d$ is a location vector, Σ_i is a $d \times d$ positive definite matrix, and $g_i(\|\mathbf{t}\|)$ is a probability density function on $\mathbb{R}^d$ for $1 \le i \le J$. If a depth function, $D(\mathbf{x}, F_i)$, is invariant under affine transformations of $\mathbf{x}$, then it is a function of $\|\Sigma_i^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_i)\|$. Using this relationship, the density, $f_i(\mathbf{x})$, can now be written as

$$f_i(\mathbf{x}) = \psi_i\{D(\mathbf{x}, F_i)\} \text{ for all } 1 \le i \le J,$$

where ψ_i is an appropriate real-valued function that depends on g_i. Now, for $1 \leq i \leq J$, it is easy to see that

$$p(i|\mathbf{x}) = p\{i|\mathbf{z}(\mathbf{x})\} = \frac{\pi_i \psi_i\{D(\mathbf{x}, F_i)\}}{\sum_{k=1}^{J} \pi_k \psi_k\{D(\mathbf{x}, F_i)\}}, \quad (1)$$

where $\mathbf{z}(\mathbf{x}) = \{z_1(\mathbf{x}), \ldots, z_J(\mathbf{x})\}^T = \{D(\mathbf{x}, F_1), \ldots, D(\mathbf{x}, F_J)\}^T$. This shows that the Bayes classifier is based on the J-dimensional vector $\mathbf{z}(\mathbf{x})$.

Spatial depth is a popular notion of data depth that was introduced and studied in Vardi & Zhang (2000) and Serfling (2002). The standardized version of *spatial depth* (SPD) of an observation, $\mathbf{x} \in \mathbb{R}^d$, with respect to a distribution function, F, on $\mathbb{R}^d$ is defined as

$$\text{SPD}(\mathbf{x}, F) = 1 - \left\| E_F[u\{\Sigma^{-1/2}(\mathbf{x} - \mathbf{X})\}] \right\|,$$

where $\mathbf{X} \sim F$ and Σ is a scatter matrix associated with the distribution F. Here $u(\cdot)$ is the multivariate sign function defined as $u(\mathbf{x}) = \|\mathbf{x}\|^{-1}\mathbf{x}$ if $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x} \in \mathbb{R}^d$, and $u(\mathbf{0}_d) = \mathbf{0}_d$, where $\mathbf{0}_d \in \mathbb{R}^d$ is the d-dimensional vector of zeros. Henceforth $\|\cdot\|$ will denote the Euclidean norm. Note that $\text{SPD}(\mathbf{x}, F_i)$ is a monotonically decreasing function of $\|\Sigma_i^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_i)\|$; see Dutta & Ghosh (2014).

To implement our approach, one can calculate the J-dimensional vector, $\mathbf{z}(\mathbf{x}_{ij})$, where $\mathbf{x}_{ij}$ is the j-th training sample in the i-th class for $1 \leq j \leq n_i$ and $1 \leq i \leq J$. These $\mathbf{z}(\mathbf{x}_{ij})$s can be viewed as realizations of the vector of covariates in a nonparametric regression model, where the response corresponds to the class label that belongs to $\{1, \ldots, J\}$. Hence, a classifier using SPD can be constructed using kernel regression; see, for example, Signorini & Jones (2004).

Our classifier is a combination of two components. The use of SPD to estimate the radial vector makes it semiparametric, while kernel density estimation is nonparametric. Performance of the nonparametric portion depends critically on bandwidth parameter values. Several techniques for choosing optimal bandwidths from the data are available in the literature; see, for example, Jones et al. (1996). However, simultaneous consideration of different levels of smoothing is expected to yield more useful information for classification. Therefore, we consider a family of density estimates, $\{\psi_{ih_i} : h_i \in H\}$, for $1 \leq i \leq J$ over a wide range of bandwidths. Multiscale methodology has emerged as a powerful exploration and visualization tool for statistical data analysis. Chaudhuri & Marron (1999) used similar methods to find significant features in regression and density estimates and Godtliebsen et al. (2002) used it to explore bivariate density estimation.

For a two-class problem and given an observation, $\mathbf{x}$, consider the following probability:

$$P_{h_1,h_2}(\mathbf{x}) = P\{p_{h_1,h_2}(1|\mathbf{x}) > p_{h_1,h_2}(2|\mathbf{x})\},$$

where

$$p_{h_1,h_2}(1|\mathbf{x}) = \frac{\pi_1 \hat{\psi}_{1h_1}\{D(\mathbf{x}, F_{1,n_1})\}}{\pi_1 \hat{\psi}_{1h_1}\{D(\mathbf{x}, F_{1,n_1})\} + \pi_2 \hat{\psi}_{2h_2}\{D(\mathbf{x}, F_{2,n_2})\}},$$

with $p_{h_1,h_2}(2|\mathbf{x}) = 1 - p_{h_1,h_2}(1|\mathbf{x})$, $\hat{\psi}_{ih_i}\{D(\mathbf{x}, F_{i,n_i})\} = \frac{1}{n_i}\sum_{k=1}^{n_i} K_{h_i}\{D(\mathbf{x}_i, F_{i,n_i}) - D(\mathbf{x}, F_{i,n_i})\}$, and the empirical distribution function, F_{i,n_i}, based on n_i observations for $i = 1, 2$. Here, the kernel function, $K_{h_i}(\cdot)$, is a 1D density function, and $h_i > 0$ is a smoothing parameter (or bandwidth) for $i = 1, 2$.

For fixed h_1 and h_2 values, the density estimates are averages of independent and identically distributed random variables. Thus, we use a normal approximation to evaluate the above mentioned probability for moderate training sample sizes. The asymptotic normality follows from the standard central limit. Using such a normal approximation with estimated means and variances, we get

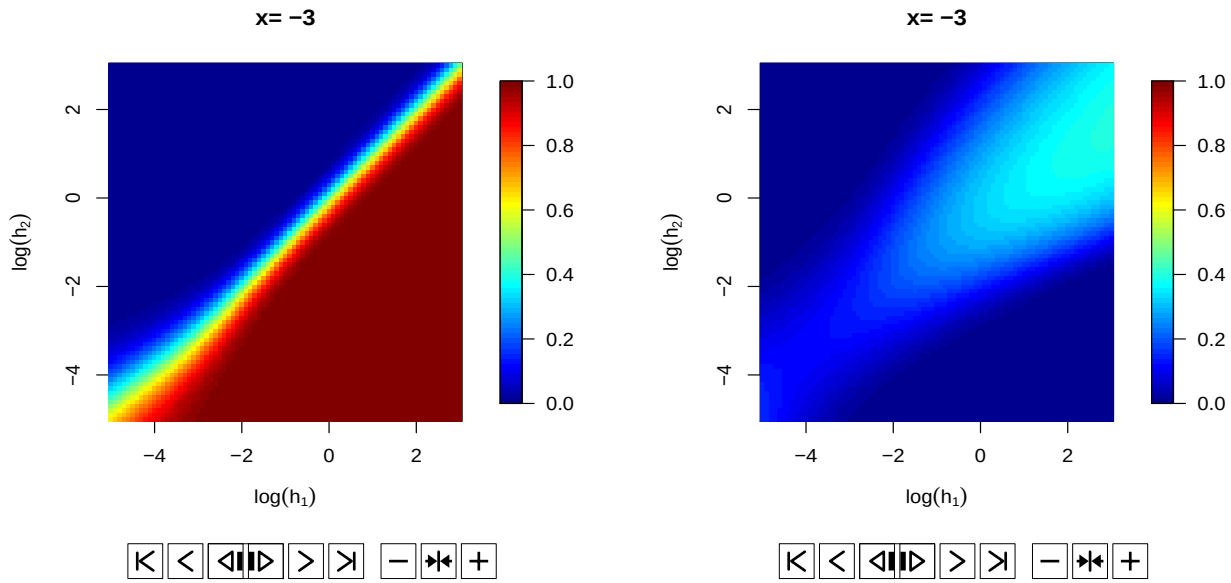
$$P_{h_1,h_2}(\mathbf{x}) \simeq \Phi\left[\frac{\pi_1 \hat{\psi}_{1h_1}\{D(\mathbf{x}, F_{1,n_1})\} - \pi_2 \hat{\psi}_{2h_2}\{D(\mathbf{x}, F_{2,n_2})\}}{\sqrt{\pi_1^2 s_{1h_1}^2/n_1 + \pi_2^2 s_{2h_2}^2/n_2}}\right],$$

where Φ is the standard normal distribution function, n_1 and n_2 are the training sample sizes for the two classes, and s_i is the variance of $\hat{\psi}_{ih_i}(\cdot)$, which can be estimated from the training sample using the sample variance of its n_i components ($i = 1, 2$). Ghosh et al. (2006) proposed this discrimination measure for multiscale analysis of kernel discriminant analysis and used static images for visualization. In this article, we use animations to explore the visualization aspect of the proposed classifier.

We first considered data for the two classes from $N_d(\mathbf{0}_d, \mathbf{I}_d)$ and $N_d(\mathbf{1}_d, \mathbf{I}_d)$, where N_d is the d-dimensional normal distribution. Here $\mathbf{0}_d$ and $\mathbf{1}_d$ are d-dimensional vectors of zeros and ones and $\mathbf{I}_d$ is the $d \times d$ identity matrix. Taking an equal number of observations from each of the two classes, we generated a training set of size 200. The sequence of test points is from the set, $\{\mathbf{x} = (x, \ldots, x)^T \in \mathbb{R}^d : x \in (-3, 3)\}$, and the set of bandwidths, $H = \{h : \log h \in (-5, 3)\}$. As a second example, the distribution of the first class is $N_d(\mathbf{0}_d, \mathbf{I}_d)$, while that for the second class is $t_{3,d}(\mathbf{0}_d, \mathbf{I}_d)$. Here $t_{m,d}(\mathbf{0}_d, \mathbf{I}_d)$ denotes the d-variate Student's t distribution with m degrees of freedom, and the vector of non-centrality parameters is $\mathbf{0}_d$. We considered both examples for $d = 2$.

The results of multiscale analysis are presented using 2D plots in Movie 2. Each frame in a figure corresponds to a specific observation that will be classified; we have a sequence of such frames to constitute this animation. The colored representation of posterior probabilities shows the natural logarithms of the bandwidths of the first and second classes plotted along the horizontal and vertical axes, respectively. Red (i.e., high posterior) indicates regions in favor of the first class, whereas blue (i.e., low posterior) points towards the second class. Color intensity varies with the magnitude of posterior probabilities, which helps us to determine the regions for strong evidence in favor of one of these two classes.

The first example (a location problem) shows an equal balance of red and blue regions in the case of $x = -3$ and $x = 3$. This is expected as none of the classes are favored when observations are far away from the center. For $x = 0.5$



Movie 2. Multiscale analysis of two-class classification: first example (left) and second example (right). Red indicates regions in favor of the first class, whereas blue points towards the second class. Color intensity varies with the magnitude of posterior probabilities.

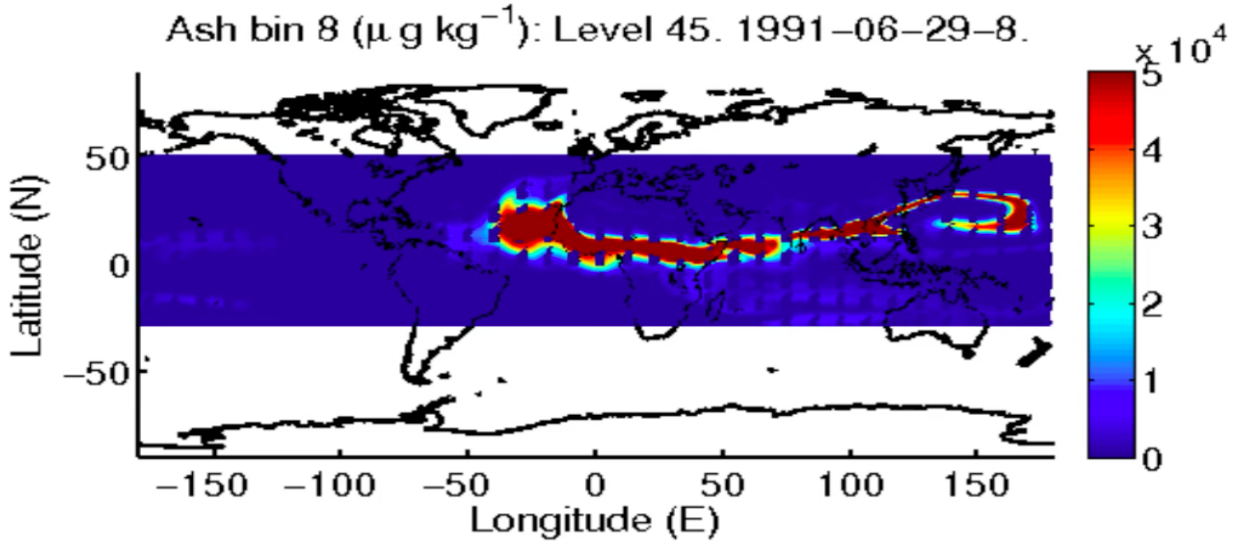
(which lies on the intersection of the two class boundaries), the evidence does not strongly favor either of the two classes; however, we see a dominance of red (respectively, blue) for x lying in the interval $(-1.5, 0.5)$ (respectively, $(0.5, 2.5)$). These regions correspond to respectively higher probability regions for the two distributions. The second example (a classification problem with difference only in shapes) illustrates that the posterior probability is in favor of the first (respectively, second) class in the $x \in (-1.25, 1.25)$ region (respectively, $x \in (-3, -2) \cup (2, 3)$). This is explained in view of the fact that the Gaussian density dominates the t_3 density only in a closed region. Movies allow for comparison of evidence in favor of either of the two classes for a sequence of points from the sample space in a continuous fashion.

4. Visualizing the effects of an explosive volcanic eruption

Global and regional climate models produce a large amount of complex-structured data. Most of the simulated variables are 4D, with both a spatial (horizontal and vertical) and a temporal component. Many different models have been developed and used extensively to address open questions in atmospheric science; for example, those related to an improved understanding on how the climate system is changing and which variables most strongly affect it, those that plan for adaptation and mitigation strategies in response to these changes, and those for weather forecast purposes. Despite their wide application, the output format is strongly model dependent; therefore, sophisticated, ad-hoc methodologies are needed to analyze them. Thus, it is essential to find an effective way to explore the large amount of data generated by these models. We propose the use of virtual reality environments to visualize complex-structured data such as those generated by global and regional models.

Here we describe the advantages of using virtual reality techniques to better understand and more easily communicate atmospheric science concepts such as the global climate effects associated with the occurrence of an explosive volcanic eruption in the tropics. We applied the Weather Research and Forecasting model coupled with chemistry (WRF-Chem) (Grell et al., 2005) to simulate volcanic ash dispersion and sulfate aerosol formation in the stratosphere as a result of Mt. Pinatubo's eruption (15 June 1991, Philippines). We focused on Mt. Pinatubo's eruption because it produced the largest volcanic cloud of the 20^{th} century and caused significant climatic effects that persisted for several years (Stenchikov et al., 2004). We applied WRF-Chem over a regional domain extending along the equatorial belt, with 400 grid cells in longitude, 100 grid cells in latitude, and a spatial resolution of 100 km. We selected Mt. Pinatubo from the Mastin & et al. (2009) database and generated the volcanic emissions using the preprocessor PREP-CHEM-SRC introduced by Freitas et al. (2011). We defined the injected mass of volcanic ash (i.e., 2 Tg) and sulfur dioxide SO_2 (20 Tg) at 20 km in the stratosphere, for an eruption duration of 8 h, and a top pressure level of 10 mbar. We specified the ash size distribution in 10 bins according to the percentage mass fractions specified in Stuefer et al. (2013) (i.e., category S3=large silicic eruptions for Pinatubo). A test case simulation for Mt. Pinatubo's eruption was run for one month, initializing the model with NCEP-DOE Reanalysis 2 data. The purpose of this study is to (i) analyze volcanic ash dispersion over the equatorial belt and (ii) simulate sulfate aerosol formation and dispersion in the stratosphere as a result of Mt. Pinatubo's eruption using WRF-Chem for the first time. We applied the RADM (Regional Acid Deposition Model) gas-phase chemical mechanism (Stockwell et al., 1990), properly modified to take SO_2 oxidation in the stratosphere into account and the MADE/SORGAM (Modal Aerosol Dynamics Model for Europe/Secondary Organic Aerosol Model) mechanism for aerosol processes (Ackermann et al., 1998; Schell et al., 2001).

In order to analyze the output from these simulations we compare traditional visualization techniques with those offered by virtual reality environments. Using traditional visualization software, we cannot investigate a 4D variable with a single image. When using a 2D plot, we can display a variable, such as volcanic ash, over the simulated domain for a specific height and time step. The frames for each simulated time step can then be collected into a movie to describe the dispersion of volcanic ash at a specific height. For example, Movie 3 was produced from Matlab frames



Movie 3. 2D dispersion of volcanic ash particles with a diameter smaller than 3 μm at a height of approximately 20 km during the month following Mt. Pinatubo's eruption.

to visualize the dispersion of volcanic ash particles with a diameter smaller than 3 μm at a height of approximately 20 km during the month following Mt. Pinatubo's eruption. This type of visualization effectively communicates why this eruption so strongly affected the earth's climate. The evolution of the eruption shows that after one month of simulation both volcanic ash and newly formed sulfate aerosol particles are globally distributed along the equatorial belt, indicating implications for the global climate. These particles remain in the atmosphere for some time, scattering solar radiation and thus cooling the earth's surface (Stenchikov et al., 2004). However, this technique does not help describe how these particles are vertically distributed in the atmosphere or how quickly they are deposited (as a function of their size). Investigating the vertical component of these particles is crucial to determine their residence time in the atmosphere and, thus, their impacts on the earth's radiative balance. A 3D plot can be used to effectively represent fields with both a spatial component and a vertical profile, although still static in time. When entering a virtual reality environment, these limitations are overcome by the possibility of interactively exploring the data set simultaneously in space and time. This environment also allows for an effective representation of the vertical profile of volcanic ash and sulfate aerosol particles and their transport at different levels in the atmosphere. Given the possibility of fully investigating a single variable, these techniques represent a unique tool for quick explorative analyses. Furthermore, multiple variables can be displayed simultaneously for comparison purposes, whereas with a 2-3D plot those features may be confounded. A 3D earth allows for an effective display of atmospheric particles of different sizes; smaller and lighter particles are located higher in the stratosphere and above larger particles (Movie 4). This demonstration was

Movie 4. 3D dispersion of volcanic ash particles during the month following Mt. Pinatubo's eruption: smaller and lighter particles are located higher in the stratosphere and above larger particles. Available under CC BY-NC-ND 4.0.

part of a spatial statistics workshop at KAUST in March 2014 and is a joint collaboration with the KAUST visualization laboratory. During this showcase, each guest was provided with 3D glasses and entered the virtual reality environment to explore the data interactively; multiple cameras in the environment were able to track head movement and to adjust perspective accordingly.

Volcanic eruption is an ideal test case to illustrate the advantages of a virtual reality environment. It is characterized by high concentrations of both ash and sulfate aerosols in specific regions of the stratosphere and near zero values elsewhere. Movie 4 also includes clips showing the workshop participants moving virtually on the earth's surface, looking at the sky above their heads, and watching the volcanic cloud pass over Saudi Arabia.

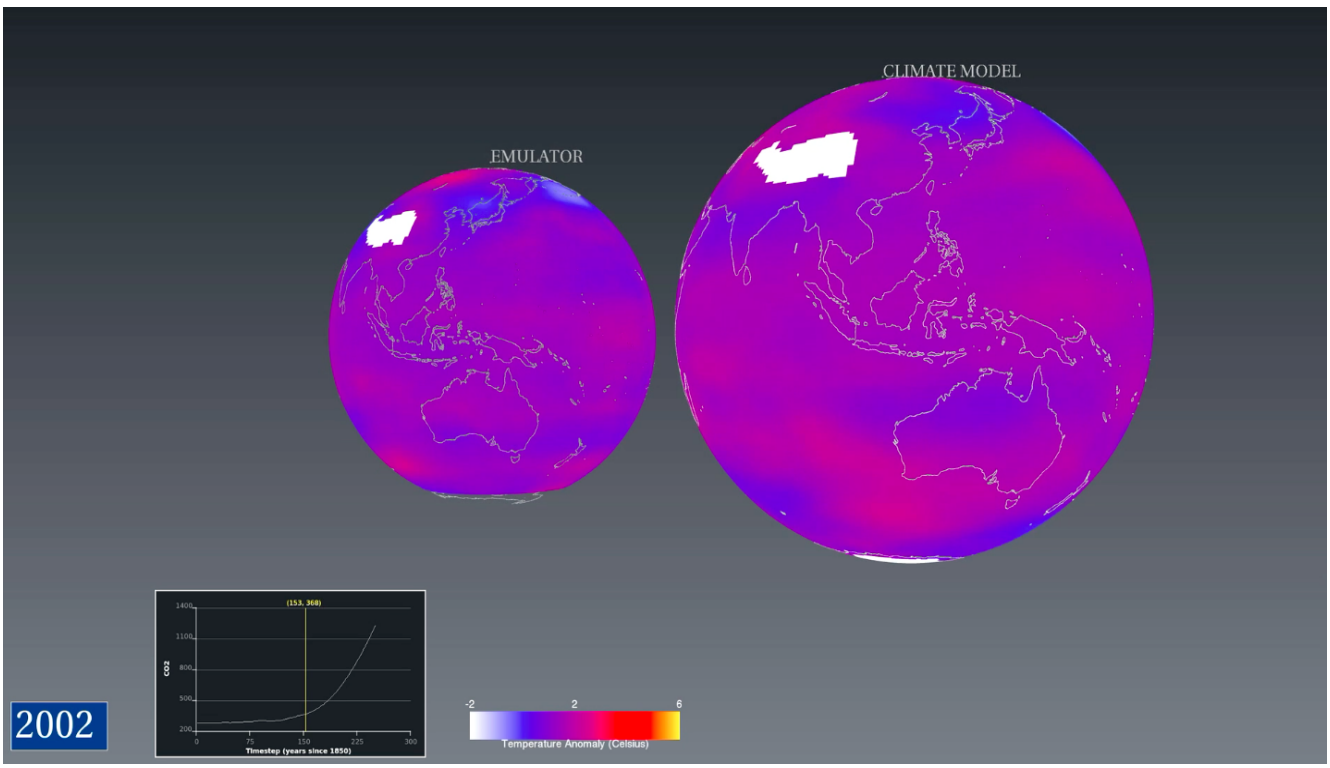
5. Visualizing and comparing global climate data

Data are commonly displayed using a spherical domain. Visualizing data, such as temperature in this context, requires a choice of a map projection that necessarily implies a distortion. 3D fields present the added challenge of representing temperatures at different heights. For extreme events, such as heat waves, floods, or volcanic eruptions, most spatial locations are uninvolved, making it possible to effectively visualize the fields with appropriate volume rendering. However, this is not possible for temperature fields where the point value at each location is of interest because

any transfer function would result in an overall opaque field at the top of the atmosphere.

In climate science, data on a 3D sphere×time domain naturally occur as output from global climate models. These models represent the state of the earth's system as coupled partial differential equations but, given the high spatio-temporal resolution, a single run could require weeks or months with state-of-the-art computational facilities. To avoid running a climate model for many input values, it is possible to define a statistical model that approximates the behavior for some input and extrapolates the behavior for different inputs. These statistical models, called emulators, are useful tools for global sensitivity analysis and parameter calibration (Sacks et al., 1989; Santner et al., 2003; Kennedy & O'Hagan, 2001; Oakley & O'Hagan, 2002; Rougier et al., 2009; O'Hagan, 2006) and impact assessment (Holden & Edwards, 2010; Castruccio & Stein, 2013; Holden et al., 2014; Castruccio et al., 2014). It is of interest to compare the emulated output with the original climate model, to understand how well the emulator approximates the original climate model and to detect possible differences. However, such a comparison is complicated by the geometry of the domain and the nature of the data. Firstly, two temperature data sets on a 3D sphere×time must be compared, but it is neither possible to effectively visualize temperature fields for different heights at the same time nor to visualize global data on a flat surface without distortion of some areas. Secondly, a sensible comparison of a climate model and an emulator cannot simply be based on a visual comparison of temperature fields for every time point, since both are weather events and the accuracy of the approximation of the emulator must be based on climate characteristics, that is, on statistical summaries.

Although some limitations are intrinsic in the nature of the data, it is possible to produce animations that allow qualitative comparison of two data sets. In this example, we build an annual 3D sphere×time temperature emulator



Movie 5. Temporal evolution of climate model output and emulation of a 3D global temperature anomaly at a given height.

of the National Center for Atmospheric Research (NCAR) CCSM4 (Gent et al., 2011) under the RCP 8.5 scenario (Van Vuuren et al., 2011) using a multi-resolution statistical model; full details can be found in Castruccio & Genton (2015). In Movie 5, two global data sets are visualized in two different boxes and show how the temperature anomaly (with respect to the first simulation year) field evolves at a given height. As time progresses, the two spheres revolve until they have rotated a full 360° by the last simulation year; rotation of the two spheres is synchronized. This movie allows the viewer to qualitatively see that the heating rate is similar for the two spheres, that it is higher at the poles, and that the spatial patterns of the two data sets do not noticeably differ. For a more formal comparison, the space-time structure of the climate model and the emulator would need to be compared; see Castruccio & Stein (2013), Castruccio & Genton (2014, 2015).

A fully immersive 3D environment allows for a comparison of multiple aspects of the data: Movie 6 shows a footage about a project on visualizing emulators in virtual reality environments. This demonstration was part of the same workshop at KAUST mentioned above. Each guest could explore the climate data and make comparisons with the emulator. This 3D environment allowed the guests to examine multiple aspects of the data simultaneously such as the surface temperature at a given height with a vertical profile at a given latitude.

Movie 6. Visualization of an emulator in a virtual reality environment. Available under CC BY-NC-ND 4.0.

6. Modeling and visualizing space-time extremes

For risk assessment, many applications require the study of the spatial or spatio-temporal extent of extreme events (e.g., in hydrology, duration of heavy storms over a certain catchment). From a statistical perspective, the max-stable paradigm provides a useful asymptotic framework for tail-dependent observations, that is, random vectors or processes with non-vanishing dependence at extreme levels. The development of theory and models for spatial extremes goes back to de Haan (1984)'s fundamental paper, which provided a useful representation of max-stable processes (slightly reformulated by Schlather, 2002), setting the basis for the construction of asymptotically justified models for extremes. Specifically, under mild conditions (see de Haan & Ferreira, 2006, Chapter 9), any max-stable process, $Z(\mathbf{x})$, indexed by $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$, with unit Fréchet margins (i.e., $\Pr\{Z(\mathbf{x}) \leq z\} = 1/z$, $z > 0$), may be represented as

$$Z(\mathbf{x}) = \sup_{i \geq 1} W_i(\mathbf{x})/P_i, \tag{2}$$

where the P_is are the points of a unit rate Poisson process on the positive half-line, and the $W_i(\mathbf{x})$s are independent copies of a positive unit mean stochastic process, $W(\mathbf{x})$, which satisfies $\mathrm{E}\{\sup_{x \in \mathcal{X}} W(\mathbf{x})\} < \infty$ in order for the supremum in (2) to be well defined. Conversely, each process with representation (2) is max-stable with unit Fréchet margins. Equation (2) may be used to construct max-stable models by choosing flexible processes, $W(\mathbf{x})$, derive finite-dimensional joint distributions (see, e.g., Davison et al., 2012), and simulate max-stable processes (Oesting et al., 2012; Dombry et al., 2013; Ribatet, 2013).

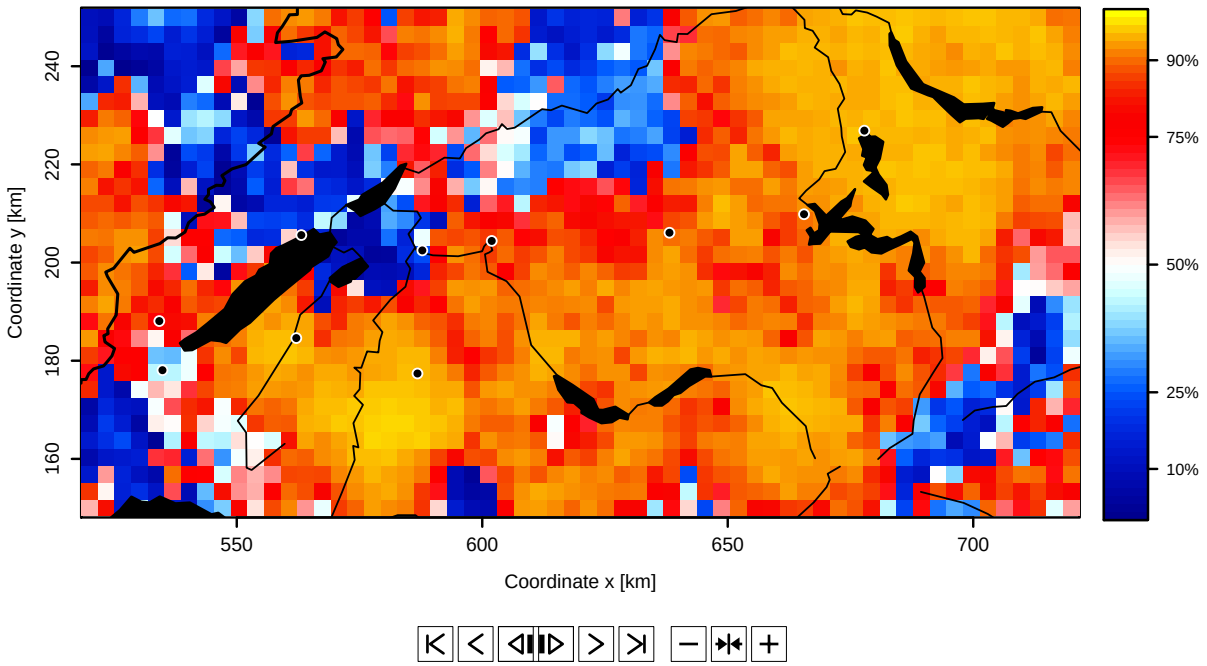
Since de Haan (1984)'s seminal paper, the literature on spatial extremes has grown rapidly and several stationary max-stable models have been proposed and fitted to real data; see, for example, Smith (1990), Padoan et al. (2010), Schlather (2002), Kabluchko et al. (2009), Reich & Shaby (2012), and the review papers by Davison et al. (2012) and Cooley et al. (2012). Applications in the space-time framework are, however, much sparser. This may be explained by additional theoretical complexities and practical difficulties, which include but are not limited to

1. the modeling of temporal dependence,
2. the non-isotropic (and perhaps also nonstationary) nature of the process of interest in the space-time domain, and the presence of non-trivial space-time interactions,
3. the validity of classic large-sample properties of widely-used composite likelihood estimators (see Padoan et al., 2010; Varin et al., 2011) when observations are time-dependent,
4. the validity of the max-stability assumption in space and time,
5. the big data problem related to the potentially very large number of space-time locations: estimation, inference, model checking, model comparison, and simulation are computationally much more complicated and demanding in space and time,
6. the visualization of space-time random fields.

Such issues have recently been partially addressed by Davis & Mikosch (2008), Davis et al. (2013a,b), and Huser & Davison (2014), though questions regarding the max-stability assumption (point 4) and useful visualization tools (point 6) have been overlooked. Next, we review the random set approach advocated by Huser & Davison (2014), used for the modeling of extremes of hourly precipitation in Switzerland, and discuss the advantage of using animations in the space-time context.

The best fit among the max-stable models considered by Huser & Davison (2014) is based on Schlather (2002) and assumes that the process, $W(\mathbf{x})$, appearing in (2) is expressed as

$$W(\mathbf{x}) \propto \max\{0, \varepsilon(\mathbf{x})\} I_{\mathcal{A}}(\mathbf{x} - \mathbf{X}), \tag{3}$$

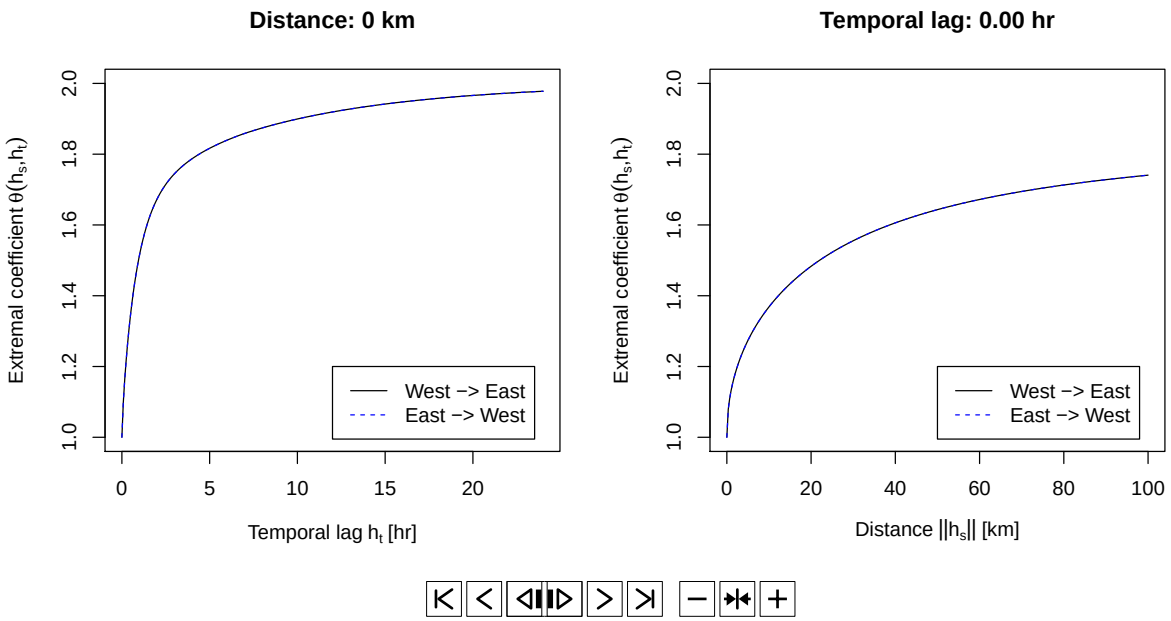


Movie 7. One realization of the space-time max-stable model (3) on a 60×30 dimensional spatial grid for 24 hours over the midland area in Switzerland. Monitoring stations are indicated with black dots, and quantiles of rainfall intensities are shown using the color scale.

where $\varepsilon(\mathbf{x})$ is a stationary standard Gaussian process on $\mathbb{R}^d$ with correlation function, $\rho(\mathbf{h})$, $\mathcal{A} \subset \mathbb{R}^d$ is a random set, $\mathbf{X}$ denotes a point of unit rate Poisson process on $\mathbb{R}^d$, and $I_{\mathcal{A}}(\cdot)$ is the indicator function on $\mathcal{A}$. Here, the dimension is $d = 3$ (plane $\times$ time), the points $\mathbf{x} = (\mathbf{s}, t) \in \mathcal{X} = \mathcal{S} \times \mathcal{T}$ have a spatial coordinate, $\mathbf{s} \in \mathcal{S}$, and a temporal coordinate, $t \in \mathcal{T}$, $\mathbf{h} = (\mathbf{h}_s, h_t)$ is a spatio-temporal separation vector, and $\mathcal{A}$ lives in the spatio-temporal domain. Thanks to the random set, which induces discontinuity in realized random fields, model (3) can be shown to be mixing, that is, it can capture independence at infinite spatio-temporal distances, a desirable feature. Furthermore, the random set, $\mathcal{A}$, can be used to mimic the underlying physical process: Huser & Davison (2014) use it to reproduce the approximate trajectories of clouds causing extreme storms across western Switzerland during the summer months. Space-time interactions are therefore controlled by two elements: long-range dependence is mainly governed by the random set, $\mathcal{A}$ (taken to be a tilted cylinder in $\mathcal{S} \times \mathcal{T}$), while small-range dependence is driven by the non-separable space-time correlation function, $\rho(\mathbf{h}_s, h_t)$ (taken from Gneiting, 2002). Model (3) was fitted to the Swiss hourly rainfall data using a pairwise likelihood, censoring observations below the 95% empirical quantile; see Huser & Davison (2014) for more detail about this approach and alternative space-time max-stable models.

A plethora of diagnostics may be used for model checking (e.g., quantile-quantile plots, extremal coefficients, and various tests), but nothing replaces visual assessment to verify whether the fitted model looks reasonable. Although it might be difficult to judge by the naked eye whether a particular model fits well, it is possible to identify pathological cases; animations are especially useful in the spatio-temporal context. As an illustration, we have simulated the fitted max-stable model (3) on a spatial grid of dimensions 60×30 for 24 hours, that is, at $43,200$ space-time locations; see Movie 7. This realization was obtained using a direct simulation method that required more than 100 Gb of memory (also indicating the need for computationally efficient simulation methods for max-stable processes). The

simulation, when viewed above the 95% quantile (yellow areas), shows that heavy storms in summer tend to move from the West to the East, as expected. The estimated 'prevailing wind direction' is 0.23 radians. Furthermore, 'wind velocity' is similar to what is observed in the real data of this study, which suggests that the fitted model does not contradict common sense. Apart from simple checks, animations allow the exploration of various features of the process of interest, such as the spatial and temporal dependence ranges. To illustrate this, Movie 8 displays fitted extremal coefficients, as a function of the temporal lag h_t and spatial distance $\|\mathbf{h}_s\|$ for two stations oriented according to the main wind direction (West $\rightarrow$ East) and two stations in the opposite direction (East $\rightarrow$ West). To be more specific, if $Z(\mathbf{s}, t)$ denotes the rainfall amount (suitably renormalized) at time t and location $\mathbf{s}$, the left panel on a single frame of Movie 8 displays estimates of the quantity $\theta(\mathbf{h}_s, h_t) \approx 2 - \Pr\{Z(\mathbf{s}+\mathbf{h}_s, t+h_t) > u \mid Z(\mathbf{s}, t) > u\}$ for large u, as a function of the temporal lag h_t. The station $\mathbf{s}+\mathbf{h}_s$ is either to the East of $\mathbf{s}$ (solid black curves) or to the West of $\mathbf{s}$ (dashed blue curves), and the different frames correspond to increasing spatial distances $\|\mathbf{h}_s\|$. Similarly, the right panel shows plots of extremal coefficients as a function of the spatial distance $\|\mathbf{h}_s\|$ for increasing temporal lags h_t. The animation reveals several interesting features of the fitted model: First, extremal dependence between the variables observed at $\mathbf{x}_1 = (\mathbf{s}_1, t_1)$ and $\mathbf{x}_2 = (\mathbf{s}_2, t_2)$ strongly depends on the relative positions of $\mathbf{s}_1$ and $\mathbf{s}_2$ (i.e., the model is not fully-symmetric, and not isotropic). Second, extremal dependence may be stronger at larger temporal lags if stations are oriented as the prevailing winds. Third, for a large temporal lag, extremal dependence is very weak, irrespective of the distance between stations. These characteristics, in agreement with the data, were already noticed by Huser & Davison (2014) but animations allow to better identify and display them. This helps to better understand the space-time dynamic of the fitted models.



Movie 8. Space-time extremal coefficients $\theta(\mathbf{h}_s, h_t)$ estimated from the fitted model in Huser & Davison (2014), plotted against the temporal lag (left) and the spatial distance (right). The different frames correspond to increasing spatial distances (left) and increasing temporal lags (right). Stations are either considered in the direction of the prevailing winds (solid black curve), or in the opposite direction (dashed blue curve).

7. Discussion

We have explored the use of visualization through animations, coined visuanimation, with a variety of examples from statistics. The main advantage of using animations is that temporal dynamics can be easily displayed rather than presenting long series of static images. We have shown that animations are useful for exploratory data analyses, as well as for model checking and comparison in complex settings such as with spatio-temporal data. More generally, embedded animations better explain features of user-developed packages for various statistics softwares.

We have illustrated the embedding of animations in the paper itself and the storage of larger movies in the online supplemental material. In the future, we expect that further progress will be made on the compression of movies, or that internet download speed will become faster, so that they can all be embedded in the paper itself. We believe that visuanimations will become the norm for many future papers in Stat.

Acknowledgement

This research was supported by King Abdullah University of Science and Technology (KAUST). The authors thank Daniel Acevedo-Feliz, April Bailey, and the KAUST Visualization Core Lab for support with the Movies 4 and 6 from the fully immersive 3D virtual reality environment.

References

Ackermann, I, Hass, H, Memmesheimer, M, Ebel, A, Binkowski, F & Shankar, U (1998), 'Modal aerosol dynamics model for europe: Development and first applications,' *Atmospheric Environment*, **32**, pp. 2981–2999.

Castruccio, S & Genton, M (2015), 'Compressing an ensemble with statistical models: An algorithm for global 3D spatio-temporal temperature,' *Technometrics*, tentatively accepted.

Castruccio, S & Genton, MG (2014), 'Beyond axial symmetry: An improved class of models for global data,' *Stat*, **3**, pp. 48–55.

Castruccio, S, McInerney, DJ, Stein, ML, Liu, F, Jacob, RJ & Moyer, EJ (2014), 'Statistical emulation of climate model projections based on precomputed GCM runs,' *Journal of Climate*, **27**, pp. 1829–1844.

Castruccio, S & Stein, ML (2013), 'Global space-time models for climate ensembles,' *Annals of Applied Statistics*, **7**, pp. 1593–1611.

Chaudhuri, P & Marron, JS (1999), 'SiZer for exploration of structures in curves,' *Journal of the American Statistical Association*, **94**, pp. 807–823.

Cooley, DS, Cisewski, J, Erhardt, RJ, Jeon, S, Mannshardt-Shamseldin, EC, Omolo, BO & Sun, Y (2012), 'A survey of spatial extremes: Measuring spatial dependence and modeling spatial effects,' *REVSTAT*, **10**, pp. 135–165.

Davis, RA, Klüppelberg, C & Steinkohl, C (2013a), 'Max-stable processes for modeling extremes observed in space and time,' *Journal of the Korean Statistical Society*, **42**, pp. 399–414.

Davis, RA, Klüppelberg, C & Steinkohl, C (2013b), 'Statistical inference for max-stable processes in space and time,' *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**, pp. 791–819.

Davis, RA & Mikosch, T (2008), 'Extreme value theory for space–time processes with heavy-tailed distributions,' *Stochastic Processes and their Applications*, **118**, pp. 560–584.

Davison, AC, Padoan, S & Ribatet, M (2012), 'Statistical modelling of spatial extremes (with discussion),' *Statistical Science*, **27**, pp. 161–186.

de Haan, L (1984), 'A spectral representation for max-stable processes,' *Annals of Probability*, **12**, pp. 1194–1204.

de Haan, L & Ferreira, A (2006), *Extreme Value Theory: An Introduction*, Springer, New York.

Dombry, C, Éyi-Minko, F & Ribatet, M (2013), 'Conditional simulation of max-stable processes,' *Biometrika*, **100**, pp. 111–124.

Dutta, S & Ghosh, AK (2014), 'Classification using localized spatial depth with multiple scales of localization,' *Under Review*.

Freitas, S, Longo, K, Alonso, M, Pirre, M, Marecal, V, Grell, G, Stockler, R, Mello, R & Sanchez Gacita, M (2011), 'Prep-chem-src-1.0: a preprocessor of trace gas and aerosol emission fields for regional and global atmospheric chemistry models,' *Geoscientific Model Development*, **4**, pp. 419–433.

Gent, P, Danabasoglu, G, Donner, L, Holland, M, Hunke, E, Jayne, S, Lawrence, D, Neale, R, Rasch, P, Vertenstein, M, Worley, P, Yang, ZL & Zhang, M (2011), 'The community climate system model version 4,' *Journal of Climate*, **24**, pp. 4973–4991.

Genton, MG, Johnson, C, Potter, K, Stenchikov, G & Sun, Y (2014), 'Surface boxplots,' *Stat*, **3**, pp. 1–11.

Ghosh, AK, Chaudhuri, P & Sengupta, D (2006), 'Classification using kernel density estimates: multiscale analysis and visualization,' *Technometrics*, **48**, pp. 120–132.

Gneiting, T (2002), 'Nonseparable, stationary covariance functions for space–time data,' *Journal of the American Statistical Association*, **97**, pp. 590–600.

Godtliebsen, F, Marron, JS & Chaudhuri, P (2002), 'Significance in scale space for bivariate density estimation,' *Journal of Computational and Graphical Statistics*, **11**, pp. 1–21.

Grell, G, Peckham, S, Schmitz, R, McKeen, S, Frost, G, Skamarock, W & Eder, B (2005), 'Fully coupled 'online' chemistry in the wrf model,' *Atmospheric Environment*, **39**, pp. 6957–6976.

Holden, PB & Edwards, NR (2010), 'Dimensionally reduced emulation of an AOGCM for application to integrated assessment modelling,' *Geophysical Research Letters*, **37**, L21707.

Holden, PB, Edwards, NR, Garthwaite, PH, Fraedrich, K, Lunkeit, F, Kirk, E, Labriet, M, Kanudia, A & Babonneau, F (2014), 'Plasim-entsem v1.0: a spatio-temporal emulator of future climate change for impacts assessment,' *Geoscientific Model Development*, **7**, pp. 433–451.

Huser, R & Davison, AC (2014), 'Space-time modelling of extreme events,' *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**, pp. 439–461.

Jones, MC, Marron, JS & Sheather, SJ (1996), 'A brief survey of bandwidth selection for density estimation,' *Journal of the American Statistical Association*, **91**, pp. 401–407.

Kabluchko, Z, Schlather, M & de Haan, L (2009), 'Stationary max-stable fields associated to negative definite functions,' *Annals of Probability*, **37**, pp. 2042–2065.

Kennedy, MC & O'Hagan, A (2001), 'Bayesian calibration of computer models,' *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**, pp. 425–464.

Liu, RY, Parelius, JM & Singh, K (1999), 'Multivariate analysis by data depth: descriptive statistics, graphics and inference,' *Annals of Statistics*, **27**, pp. 783–858, with discussion and a rejoinder by Liu and Singh.

López-Pintado, S & Romo, J (2009), 'On the concept of depth for functional data,' *Journal of the American Statistical Association*, **104**, pp. 718–734.

Mastin, LG & et al. (2009), 'A multidisciplinary effort to assign realistic source parameters to models of volcanic ash-cloud transport and dispersion during eruptions,' *Journal of Volcanology and Geothermal Research*, **186**, pp. 10–21.

Oakley, O & O'Hagan, A (2002), 'Bayesian inference for the uncertainty distribution of computer model outputs,' *Biometrika*, **89**, pp. 769–784.

Oesting, M, Kabluchko, Z & Schlather, M (2012), 'Simulation of Brown–Resnick processes,' *Extremes*, **15**, pp. 89–107.

O'Hagan, A (2006), 'Bayesian analysis of computer code output: A tutorial,' *Reliability Engineering and System Safety*, **91**, pp. 1290–1300.

Padoan, SA, Ribatet, M & Sisson, SA (2010), 'Likelihood-based inference for max-stable processes,' *Journal of the American Statistical Association*, **105**, pp. 263–277.

Reich, BJ & Shaby, BA (2012), 'A hierarchical max-stable spatial model for extreme precipitation,' *Annals of Applied Statistics*, **6**, pp. 1430–1451.

Ribatet, M (2013), 'Spatial extremes: Max-stable processes at work,' *Journal de la Société Française de Statistique*, **154**, pp. 156–177.

Rougier, J, Sexton, D, Murphy, J & Stainforth, D (2009), 'Analyzing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments,' *Journal of Climate*, **22**, pp. 3540–3557.

Rousseeuw, PJ, Ruts, I & Tukey, JW (1999), 'The bagplot: A bivariate boxplot,' *The American Statistician*, **53**, pp. 382–387.

Sacks, J, Welch, W, Mitchell, T & Wynn, H (1989), 'Design and analysis of computer experiments,' *Statistical Science*, **4**, pp. 409–423.

Santner, TJ, B., W & W., N (2003), *The Design and Analysis of Computer Experiments*, Springer-Verlag, New York.

Schell, B, Ackermann, I, Hass, H, Binkowski, F & Ebel, A (2001), 'Modeling the formation of secondary organic aerosol within a comprehensive air quality model system,' *Journal of Geophysical Research: Atmospheres*, **106**, pp. 28275–28293.

Schlather, M (2002), 'Models for stationary max-stable random fields,' *Extremes*, **5**, pp. 33–44.

Serfling, R (2002), 'A depth function and a scale curve based on spatial quantiles,' in *Statistical data analysis based on the L_1-norm and related methods (Neuchâtel, 2002)*, Birkhäuser, Basel, Stat. Ind. Technol., pp. 25–38.

Signorini, DF & Jones, MC (2004), 'Kernel estimators for univariate binary regression,' *Journal of the American Statistical Association*, **99**, pp. 119–126.

Smith, RL (1990), 'Max-stable processes and spatial extremes,' Unpublished.

Stenchikov, G, Hamilton, K, Robock, A, Ramaswamy, V & Schwarzkopf, M (2004), 'Arctic oscillation response to the 1991 pinatubo eruption in the skyhi general circulation model with a realistic quasi-biennial oscillation,' *Journal of Geophysical Research: Atmospheres*, **109 (D3)**, pp. 1–13.

Stockwell, WR, Middleton, P, Chang, J & Tang, X (1990), 'The second generation regional acid deposition model chemical mechanism for regional air quality modeling,' *Journal of Geophysical Research: Atmospheres*, **95 (D10)**, pp. 16343–16367.

Stuefer, M, Freitas, S, Grell, G, Webley, P, Peckham, S, S.and McKeen & Egan, S (2013), 'Inclusion of ash and so_2 emissions from volcanic eruptions in wrf-chem: development and some applications,' *Geoscientific Model Development*, **6**, pp. 457–468.

Sun, Y & Genton, MG (2011), 'Functional boxplots,' *Journal of Computational and Graphical Statistics*, **20**, pp. 316–334.

Sun, Y & Genton, MG (2012), 'Adjusted functional boxplots for spatio-temporal data visualization and outlier detection,' *Environmetrics*, **23**, pp. 54–64.

Sun, Y, Genton, MG & Nychka, D (2012), 'Exact fast computation of band depth for large functional datasets: How quickly can one million curves be ranked?' *Stat*, **1**, pp. 68–74.

Tukey, JW (1975), 'Mathematics and the picturing of data,' in James, RD (ed.), *Proceedings of the International Congress of Mathematicians*, Canadian Mathematical Society, Vancouver, pp. 523– 531.

Van Vuuren, D, Edmonds, J, Kainuma, M, Riahi, K, Thomson, A, Hibbard, K, Hurtt, G, Kram, T, Krey, V, Lamarque, JF, Masui, T, Meinshausen, M, Nakicenovic, N, Smith, S & Rose, S (2011), 'The representative concentration pathways: an overview,' *Climatic Change*, **109**, pp. 5–31.

Vardi, Y & Zhang, CH (2000), 'The multivariate L_1-median and associated data depth,' *Proceedings of the National Academy of Sciences of the United States of America*, **97**, pp. 1423–1426 (electronic).

Varin, C, Reid, N & Firth, D (2011), 'An overview of composite likelihood methods,' *Statistica Sinica*, **21**, pp. 5–42.

Zuo, Y & Serfling, R (2000), 'General notions of statistical depth function,' *Annals of Statistics*, **28**, pp. 461–482.

Supporting Information

Additional supporting information (Movies 1,3,4,5,6) may be found in the online version of this article at the publisher's web site. A link has been implemented from the pdf file of the paper to these movies which need to be located in the same directory as the pdf file. In order to play these movies in the pdf file on a Windows computer with Adobe Reader, QuickTime Player for Windows must be downloaded and installed. Further, QuickTime Player must be set as Preferred Media Player in Edit -> Preferences -> Multimedia (legacy). Finally all the movie files must be added using Tools -> Send File: -> + Add File. With Adobe Pro in Windows and with Apple and Linux devices the movies can be played directly. The other Movies 2,7,8 are directly embedded in the pdf file.