EDITOR: Amit Agrawal, amit.agrawal@me.com
GUEST EDITOR: Francesca Samsel, fsamsel@tacc.utexas.edu

DEPARTMENT: TOOLS AND PRODUCTS

Visualization of Climate Science Simulation Data

Niklas Röber and Michael Böttinger, *German Climate Computing Center, Hamburg, Germany*

Bjorn Stevens, *Max Planck Institute for Meteorology, Hamburg, Germany*

Climate simulations belong to the most data-intensive scientific disciplines and are—in relation to one of humankind's largest challenges, i.e., facing anthropogenic climate change—ever more important. Not only are the outputs generated by current models increasing in size, due to an increase in resolution and the use of ensembles, but the complexity is also rising as a result of maturing models that are able to better describe the intricacies of our climate system. This article focuses on developments and trends in the scientific workflow for the analysis and visualization of climate simulation data, as well as on changes in the visualization techniques and tools that are available.

Climate research has come a long way, from models and observations that coarsely describe Earth's atmosphere and ocean in the past, to dense satellite data and km scale global simulations that explicitly resolve clouds and precipitation today. The simulations that are carried out at DKRZ, the German Climate Computing Center in Hamburg, Germany, span a wide range from regional to global, from short to very long, and from low to very high resolution.

A commonality of all simulations is the multivariate nature of the data that is written out, with a multitude of time-varying 2-D and 3-D variables describing the states of the atmosphere, ocean, cryosphere, and biosphere. The development of expressive, yet comprehensive, visualizations by combining information from several different variables to unlock correlations and trends can sometimes be considered more an art than science. Over the last decades, the output generated by climate models has grown exponentially. This is essentially due to higher spatial and temporal resolutions. The introduction of ensemble techniques, as well as additional processes being incorporated into the models and output as data, also add to the total sum. One tool that is particularly good at data reduction is *visualization*, yet the data we are working with these days is so detailed that we often do not see the forest for the trees.

Large data comes in different forms, requiring different solutions each to be applied in order to *visualize* the information within. While a few years back 80 km global simulations had been considered state of the art, producing about 2 GB of data per simulated day, nowadays we are working on 1 km global simulations requiring to store at least 8 TB of data per simulated day with the same number of time steps[1, 2] (see Figure 1). Yet at the same time, the analysis of small-scale processes that can now be resolved demand a particularly higher frequency output in order to be visualized properly, thus further increasing the amount of data by a large factor. As storage capacities are tight, one needs to carefully choose what to store and what not.

Another form of large data emanates from very long simulations, such as within the German Government funded project PALMOD, which simulates the climate from the last interglacial to the Anthropocene, i.e., a complete glacial cycle of about 120 k years.[3] While the spatial resolution of the model is low, data extraction and the visualization of many thousand time steps, and especially finding key events in this long dataset, is not a trivial task (see Figure 2).

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Digital Object Identifier 10.1109/MCG.2020.3043987
Date of current version 12 January 2021.

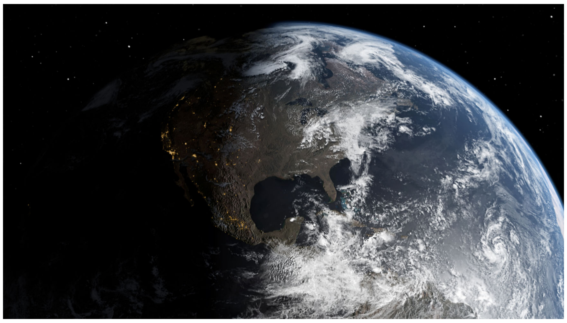


FIGURE 1. Photo-realistic visualization of a global ICON atmosphere simulation at 2.5 km resolution. Clouds are a 2-D composition of vertically integrated liquid cloud water and ice, and rendered using a physically based rendering (PBR) approach.

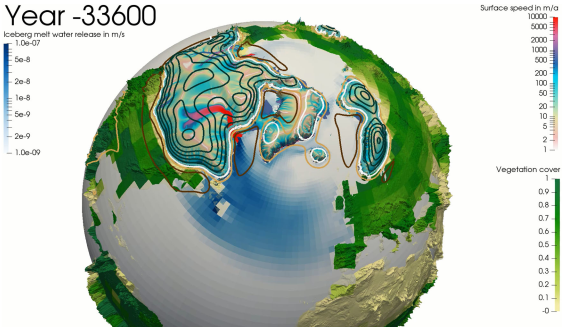


FIGURE 2. Depicted in this visualization from a paleoclimate simulation is a so-called *Heinrich* event, in which large quantities of ice flow into the Atlantic and weaken the circulation, with a drastic effect on the climate in Europe. (Source: Florian Ziemen, DKRZ; used with permission.).

A third form of large data is generated by ensemble simulations, in which a climate model is started several times with varied initial conditions. By sampling the uncertainty in the initial conditions, the ensemble provides means to derive additional statistical information such as the internal climate variability and hence the uncertainty or robustness of a result. The Max Planck Institute for Meteorology (MPI-M) grand ensemble[4] (MPI-GE) is to date the largest climate projection ensemble of a single state-of-the-art comprehensive earth system model; it consists of 100 members each for three CMIP5 scenarios, a 2000 year preindustrial control run, a simulation of the historical past from 1850 to 1995, and a 150 year generic greenhouse gas increase experiment (see Figure 3).

Of course, not all simulations that are performed at DKRZ are very large; in terms of raw numbers, the majority are not. Our archive already holds about 140 PB of climate simulation data for long-term storage. With increasing computing capacity and progressively complex simulations, this number will rise at an even higher rate. The only question that remains is: *How are we coping with this huge amount of data in the near future?* This article tries to answer this question, by discussing the climate *Visualization Workflow*, as well as the *Visualization Techniques* employed, thereby illuminating both aspects with respect to past, current, and future approaches.

VISUALIZATION WORKFLOW

Besides a supercomputer and a large parallel file system, DKRZ also hosts a dedicated visualization cluster as an integral part of the HPC system. The visualization nodes are equipped with GPUs, and are configured with X, VirtualGL, and VNC to facilitate remote access and rendering. As these GPU nodes share the file system with the supercomputer, no data has to leave the data center, and the entire workflow stays *in-house*.

Our primary applications for postprocessing, analysis, and visualization are the climate data operators CDO (https://code.mpimet.mpg.de/projects/cdo), the netCDF operators NCO, NCL (https://www.ncl.ucar.edu), Python/matplotlib, among of course other libraries such as DASK and XArray, but also ParaView,[5] Met3D,[6] and VaPOR.[7] In our work, we try to employ open-source software as much as possible, and also foster the open-source community through collaborations and user training. For a broader overview on tools and developments for data visualization in this field, we refer to the review by Rautenhaus.[8]

The classic workflow in climate science is the one of posthoc data analysis and visualization: During the simulation, and depending on storage capacities available, as much data are written out as possible. Later, the data are processed, analyzed, and visualized, and some of the data may also get archived in tape libraries for long-term storage. However, the widening of the gap between computational performance and I/O bandwidth invalidates this concept, even more so for larger simulations, as only fractions of the data can be stored for later posthoc analyses. Especially for very data-intensive experiments, scientists must think ahead and formulate the research questions they would like to get answered with the simulations beforehand, as well as carefully plan which data to store and which to not. This is very similar to the data acquisition process in measurement campaigns, and needs now to be transitioned to the simulation planning.

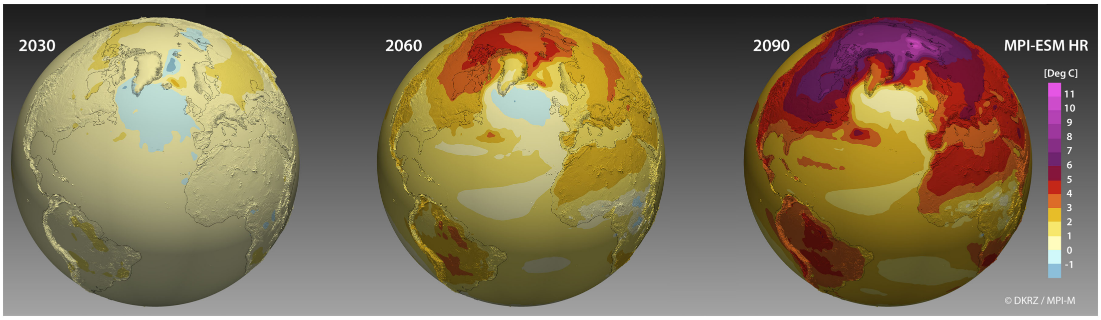


FIGURE 3. Simulations carried out in the context of the Coupled Model Intercomparison Project 6 (CMIP6) for the upcoming sixth IPCC report are a good example for long-term ensemble climate projections. The visualization shows projected temperature changes for 2030, 2060, and 2090 for the pessimistic scenario SSP585.

Especially for these big flagship simulations, the workflow is about to change. With the advent of ICON—the ICOsahedral Nonhydrostatic model[9] that is jointly developed by the MPI-M, the German Weather Service (DWD), and DKRZ—the necessity arose to read and visualize these ICON datasets on the original unstructured grid. As ParaView is open source and being used by many institutions from various scientific disciplines, we decided to extend ParaView with a customized reader to directly import and visualize ICON netCDF/grib compressed data. Our reader employs the climate data interface CDI (https://code.mpimet.mpg.de/projects/cdi), also developed by MPI-M, as underlying netCDF library, and is not only able to read all three ICON grids (triangles, quads, and hexagons), but also the output of other models such as IFS and DYNAMICO. Currently, it is being further extended to support a variety of different models employed in climate and weather research.

Jupyter Notebooks have established themselves as indispensable assets for the analysis and visualization of data, not just within climate science. They are easy to use and share, and can be employed for preprocessing and postprocessing, data analysis and visualization, as well as to document the entire process. In general, web-based visualizations are gaining more momentum, not only for communication and public outreach, but also as a tool for scientists to analyze their data using various web-based services. Beneficial thereby is that these services run on servers that are also connected to the large file system, accessing all data *in-house* without any data movement.

Large Data Visualization

One barrier that poses real challenges is the one put up by storage constraints of the HPC system itself, but also by limitations of the human mental perception and processing capabilities. The performance limitations of the human eye are estimated to be around 6 Mb/s, whereas the throughput of the visual sensory system, i.e., perception and cognition, is considerably less. The optical nerve throughput is in the order of 20 kB/s, therefore, looking at petabytes of visualized data will not get us any further. We need to reduce the amount of data tremendously, and ideally steer the scientists' attention to areas and time steps that are relevant, i.e., which show unexpected changes, outliers, or undiscovered phenomena.

For the computational part of large data visualization, two primary concepts have evolved: *in situ* visualization as well as progressive data storage, access, and rendering. Both have their respective advantages and limitations, but help us to manage the output that is generated by extremely large and high-resolution experiments.

In situ visualization processes and visualizes the data as it is being computed. This has the advantage that both the I/O and the time to solution are drastically reduced. In our experience, developing and experimenting with an *in situ* adaptor for our Fortran ICON model based on ParaView/Catalyst, and running it together with the simulation in a tightly coupled setup, the performance overhead required for the processing and visualization is almost negligible[10]. At the moment, there are a lot of ongoing developments and community efforts aiming in the same direction. A number of those support web-based solutions, such as the Cinema (https://cinemascience.github.io) extension of Catalyst, which implements an image-based data visualization, but also allows data artifacts, and in its newest addition also deferred rendering.

Nevertheless, one of the major advantages of *in situ* visualization is at the same time also its biggest flaw: Only a fraction of the simulated data is stored, either as reduced data and/or rendered as images. If the structure or feature that a scientist is interested in is not contained in the data stored, or visible in one of the visualizations, the simulation may need to be rerun with revised parameters. But as the degree of parallelism of HPC systems grows, this problem might mitigate in the future, and it will be more favorable to rerun simulations over storing all the data. Future workflows might also expose steering as an option, in which the user controls parameters of the simulation directly from within the visualization application. We are currently taking our first steps to experiment with this feature.

Progressive data storage and rendering can help remedy some of the issues associated with *in situ* visualization, although it does not solve the I/O problem, as still almost the same amount of data needs to be stored to disk for a posthoc analysis. Here, the data are decomposed, often by using wavelets, to reconstruct a level-of-detail tree of the data, ideally also applying lossy data compression at the same time[11]. Later, the part with the lowest resolution is loaded and visualized first. The user can either load new details on demand or load them automatically by zooming into the data. Important thereby is to preserve the *interesting* features in the lower resolution versions, so that one knows where to zoom in to further explore the data. A visualization package that excels at this is VaPOR, which is developed by NCAR, the National Center for Atmospheric Research in Boulder, CO, USA. In a collaboration with NCAR, we work on data decomposition and visualization techniques for unstructured datasets to allow our scientists to also utilize the progressive data access and rendering that is offered by VaPOR.[7]

Visualization Automation

Artificial neural networks and automated algorithms are increasingly attracting the attention of the data analysis and visualization community, but have also caught the attention of various domains that either deal with and/or generate large datasets. At DKRZ, we are experimenting with AI workflows to automatize not only the analysis and visualization but also to improve the efficiency of the simulation itself. On the analysis side, a neural network can be used to supervise the progression of a simulation, and may stop it in case of errors to conserve computational resources. It can also be used to detect outliers and *interesting events* and directly steer the scientists' attention, as well as to learn to differentiate patterns in the data, such as cloud types, and track their development and evolution over time. We also experiment with the possibility to train a neural network online, by routing the training data directly over the *in situ* Catalyst adaptor to an AI training framework.

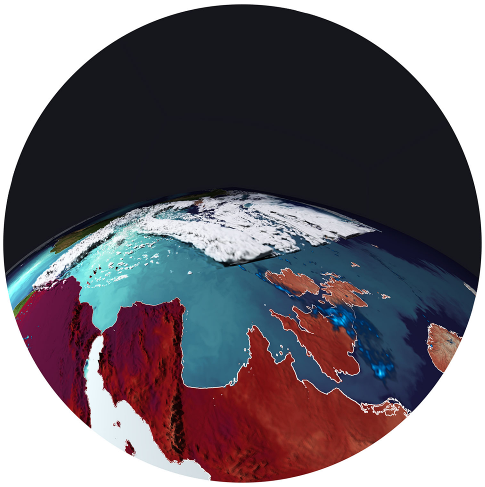


FIGURE 4. Excerpt from our DYAMOND++ VR film (https://youtu.be/5Y_oDaFRLaI), looking from central Europe over the Atlantic and onto an Icelandic low. The clouds are a 3-D composition of liquid cloud water and cloud ice, cut in half to visualize a large-scale precipitation band (blue) over the U.K. The ocean's surface shows salinity, whereas the land surface depicts temperature.

In general, we hope that these algorithms, along with an automation of the visualization pipeline, can help us to tame the data output by high-resolution models and ensemble simulations.

VISUALIZATION TECHNIQUES

Climate scientists are still fond of 2-D data visualization, as are physicians when looking at their CT and MRT scans. Some still prefer printing all their variables onto paper and comparing the images side by side. However, over the past decade, new visualization techniques enabling an interactive visual data analysis have emerged and quickly transitioned from scientific visualization research into publicly and free available interactive visualization tools.

At DKRZ, visualizations are created for a number of reasons, i.e., exploration, verification, and communication of simulation results. Most simulations have scripts running along that create standard 2-D visualizations, which are automatically generated and published onto a web server to track the plausibility of the ongoing experiment. Later, scientists explore the data posthoc using various and very common 2-D and/or 3-D data visualization techniques in order to find or verify structures, features, and correlations. Statistical data analysis plays a big role in

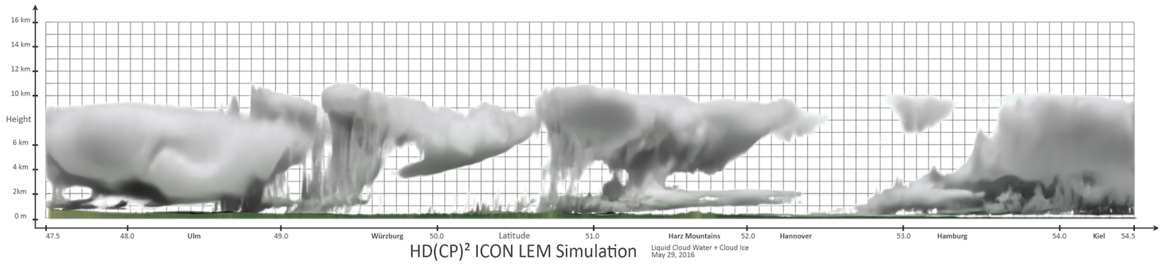


FIGURE 5. Visualization of 3-D liquid cloud water and cloud ice as vertical cross section through Germany along 10 East using volume rendering with path tracing (OSPRay) enabled. Visible are large Cumulonimbus and Nimbo Stratus cloud systems, as well as lower clouds.

climate science. A lot of work is spent on data postprocessing, i.e., to derive monthly or yearly means, to compute differences, or to filter and subset the data. For 2-D and $2\frac{1}{2}$D visualizations, scripting-based tools such as NCL and Python are employed, whereas ParaView, VaPOR, and Met3D are used for 3-D data visualization. Interactive data analysis and visualization are especially important for scientists browsing their data in an intuitive, natural way, looking for new structures and/or correlations, but also for debugging the climate model though data inspection. Animations and stills are created to communicate and discuss the results with fellow scientists, as well as for public outreach and to inform the general public (see Figure 4).

A universal difficulty thereby is the multivariate nature of the data, which further increases with the visualization of ensembles, as also the certainty/uncertainty of the data needs to be visually expressed. While in the past, the visualization of climate data was generally centered around a representation of one or two variables to show their development over time, visualizations now tend to be more complex. This is partially due to a maturing of the simulation models, which are now able to describe more (small scale) processes, but is also due to advances in visualization science. In a balance between functionality and aesthetics, one needs to find a way that the variables that are shown complement each other in telling the underlying story of science. The clouds in Figure 4, for example, are cut in half, yet by observing the blue precipitation band, the surface temperature, as well as the salinity, one can estimate the stretching of the clouds over Europe.

Improving Visual Quality

For ParaView, the integration and addition of new features has increased exponentially lately, benefiting the entire visualization community. These additions also include advances to lift the visual quality of the renderings by a number of magnitudes. Two raytracing backends are now being shipped with ParaView: OSPRay from INTEL,[12] Figure 5, and OptiX from NVIDIA,[13] Figure 6, among other great improvements, such as a PBR, the use of materials and shaders, as well as the possibility to directly render to VR headsets and produce visualizations for large dome displays (see Figure 4). Those qualitative improvements directly benefit the perception of increasingly complex simulations through better depth perception and spatial recognition.

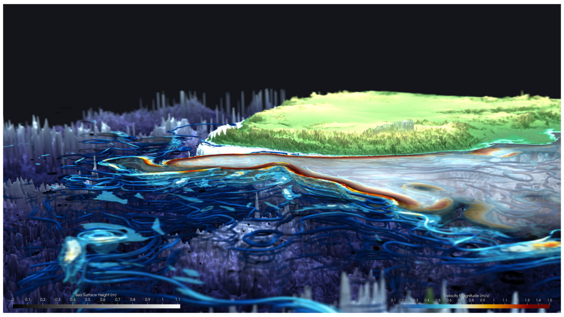


FIGURE 6. Visualization of the Agulhas current at the southern tip of Africa. Visualized are eddies using streamlines depicting the ocean's velocity, as well as sea surface height as semitransparent black and white overlay, rendered using raytracing (OptiX) and with the depth of field to steer the viewers' attention.

Additionally, raytracing outperforms OpenGL rasterization for very large triangle counts, depending on screen resolution and settings due to early ray termination, making it the ideal choice for the visualization of extremely large simulation datasets.

Feature Detection

In order to make the invisible visible and accessible to the user, the main structures and features of a dataset need to be revealed. A visualization of the raw simulation output is only feasible for smaller simulations; in larger simulations, users are literally drowned by data, and unable to perceive the entire visualization, while not getting lost in the details. In these cases it is helpful to extract and visualize the important phenomena, for example, coherent Lagrangian structures, to characterize the primary flow field dynamics. Other examples in climate science are an eddy detection for ocean and a cloud classification for high resolved atmospheric data. The eddy detection can be used for an eddy census and a quantitative visualization of eddies, revealing a number of insights to oceanographers, whereas cloud classification helps atmospheric scientists to quickly find specific cloud types and transitions, as well as to form a better understanding of cloud building and precipitation processes.

Related to this is of course the area of topology: A topological analysis can reveal structures in the data and correlations between variables and/or processes that would otherwise be hidden. A good example is fiber surface, an extension of the classic iso-surface to many variables. 3-D fiber surfaces are constructed by intersecting the iso-surfaces of the different variables used; indicating the volume, in which the thresholds of the respective iso-surfaces are met or surpassed. For the analysis of climate simulation data, this can be used to visualize, for example, relations between several hydrometeorological quantities and the wind field, in order to show the updraft of moisture as the primary *power supply* of a cumulonimbus cloud system in a high-resolution atmospherical simulation.[14]

THE FUTURE

While the prediction of the future is not easy, an extrapolation from the past is nonetheless possible: Climate scientists will continue to crank up the resolution of the models, thereby resolving and including more (small scale) processes into the simulation, as well as conduct more ensembles. Yet also the simulations will hit a barrier soon and changes within the simulation workflow itself are not too far away in the future.

With growing data sizes, we need to pay more attention to both data analysis and visualization, as this is what will be used in the end. No simulation is run just because it can be done, but because one wants to gain knowledge and learn from what is hidden within the data. With growing data sizes, we need to automate the analysis pipeline and develop tools that assist us in harvesting this information without getting lost in the noise. We need tools that are able to steer the scientists' attention to specific areas and time steps of interest, even though these areas and time steps may vary from scientist to scientist depending on the underlying research question.

REFERENCES

1. B. Stevens *et al.*, "DYAMOND: The dynamics of the atmospheric general circulation modeled on non-hydrostatic domains," *Prog. Earth Planet Sci.*, vol. 6, 2019, Art. no. 61.
2. B. Stevens *et al.*, "Large-eddy and storm resolving models for climate prediction: The added value for clouds and precipitation," *J. Meteorol. Soc. Jpn.*, vol. 98, no. 2, pp. 395–435, 2020.
3. M. Latif, M. Claussen, M. Schulz, and T. Brucher, "Comprehensive earth system models of the last glacial cycle," *Eos*, vol. 97, 2016.
4. N. Maher *et al.*, "The Max Planck Institute Grand Ensemble: Enabling the exploration of climate system variability," *J. Adv. Model. Earth Syst.*, vol. 11, no. 7, pp. 2050–2069, 2019.
5. U. Ayachit, *The ParaView Guide: A Parallel Visualization Application*. Clifton Park, NY, USA: Kitware, 2015.
6. M. Rautenhaus, M. Kern, A. Schafler, and R. Westermann, "Three-dimensional visualization of ensemble weather forecasts—Part 1: The visualization tool Met.3D (version 1.0)," *Geosci. Model Develop.*, vol. 8, pp. 2329–2353, 2015.
7. S. Li *et al.*, "VAPOR: A visualization package tailored to analyze simulation data in earth system science," *Atmosphere*, vol. 10, no. 9, 2019, Art. no. 488.
8. M. Rautenhaus *et al.*, "Visualization in meteorology—A survey of techniques and tools for data analysis tasks," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 12, pp. 3268–3296, Dec. 2018.
9. M. A. Giorgetta *et al.*, "ICON-A, the atmosphere component of the ICON Earth system model: I. Model description," *J. Adv. Model. Earth Syst.*, vol. 10, pp. 1613-1637, 2018.
10. N. Rober and J. F. Engels, "In-Situ processing in climate science," in *Proc. 4th Workshop In Situ Vis.*, 2019, pp. 612–622.
11. M. I. Jubair *et al.*, "Icosahedral maps for a multiresolution representation of earth data," in *Proc. Conf. Vision, Model. Vis.*, Bayreuth, Germany, 2016.

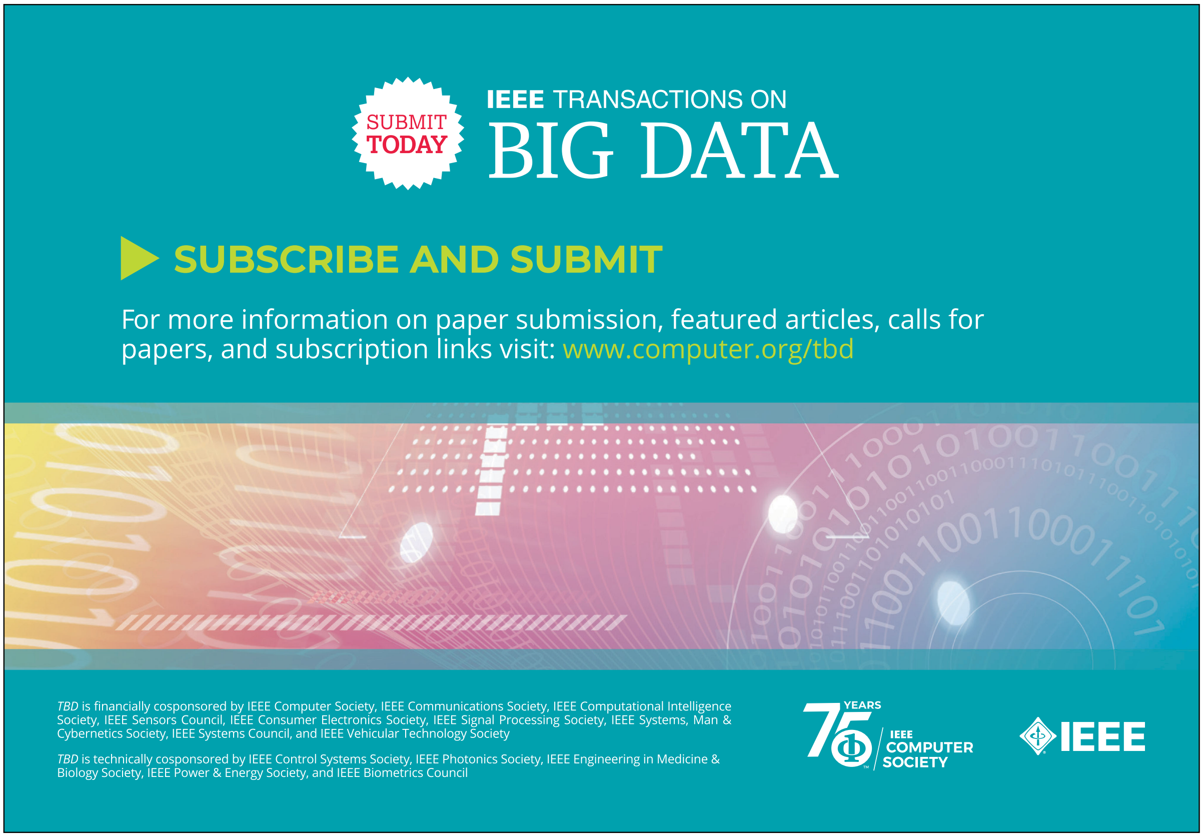
12. I. Wald *et al.*, "OSPRay—A CPU ray tracing framework for scientific visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 931–940, Jan. 2017.
13. S. Parker *et al.*, "OptiX: A general purpose ray tracing engine," *ACM Trans. Graph.*, vol. 29, no. 4, 2010.
14. C. Blecha *et al.*, "Fiber surfaces for many variables," *Comput. Graph. Forum*, vol. 39, no. 3, pp. 317–329, 2020.

NIKLAS RÖBER is a Climate Visualization Expert with the German Climate Computing Center (DKRZ), Hamburg, Germany, which he joined in 2009. His research interests include visualization and analysis of simulation data, especially the visualization of extremely large and unstructured climate datasets. He received the Ph.D. degree in computer science and an MBA. He is the corresponding author of this article. Contact him at roeber@dkrz.de.

MICHAEL BÖTTINGER leads the visualization and public relations group and his research is focused on scientific visualization of climate model data. He received the Diploma in geophysics and started as a Scientist with the Max Planck Institute for Meteorology, Hamburg, Germany, in 1988, before joining the German Climate Computing Center (DKRZ), Hamburg, Germany, in 1999. Contact him at boettinger@dkrz.de.

BJORN STEVENS is a Director of the Max Planck Institute for Meteorology, Hamburg, Germany, where he leads the Atmosphere in the Earth System Department. He is also a Professor with the University of Hamburg, Hamburg, Germany. His main interest is in the way atmospheric water—particularly in the form of clouds—shapes the climate. Contact him at bjorn.stevens@mpimet.mpg.de.

Contact department editor Amit Agrawal at amit.agrawal@me.com.



SUBMIT TODAY

IEEE TRANSACTIONS ON BIG DATA

SUBSCRIBE AND SUBMIT

For more information on paper submission, featured articles, calls for papers, and subscription links visit: www.computer.org/tbd

TBD is financially cosponsored by IEEE Computer Society, IEEE Communications Society, IEEE Computational Intelligence Society, IEEE Sensors Council, IEEE Consumer Electronics Society, IEEE Signal Processing Society, IEEE Systems, Man & Cybernetics Society, IEEE Systems Council, and IEEE Vehicular Technology Society

TBD is technically cosponsored by IEEE Control Systems Society, IEEE Photonics Society, IEEE Engineering in Medicine & Biology Society, IEEE Power & Energy Society, and IEEE Biometrics Council

75 YEARS IEEE COMPUTER SOCIETY

IEEE