Managing Large-scale Atmospheric and Oceanic Climate Data for Efficient Analysis and On-the-fly Interactive Visualization

Aashish Panta[1], Giorgio Scorzelli[1], Amy Gooch[1], Valerio Pascucci[1], and Huikyo Lee[2]

[1]University of Utah
[2]JPL/NASA/Caltech

November 23, 2024

Abstract

Managing vast volumes of climate data, often reaching into terabytes and petabytes, presents significant challenges in terms of storage, accessibility, efficient analysis, and on-the-fly interactive visualization. Traditional data handling techniques are increasingly inadequate for the massive atmospheric and oceanic data generated by modern climate research. We tackled these challenges by reorganizing the native data layout to optimize access and processing, implementing advanced visualization algorithms like OpenVisus for real-time interactive exploration, and extracting comprehensive metadata for all available fields to improve data discoverability and usability. Our work utilized extensive datasets, including downscaled projections of various climate variables and high-resolution ocean simulations from NEX GDDP CMIP6 and NASA DYAMOND datasets. By transforming the data into progressive, streaming-capable formats and incorporating ARCO (Analysis Ready, Cloud Optimized) features before moving them to the cloud, we ensured that the data is highly accessible and efficient for analysis, while allowing direct access to data subsets in the cloud. The direct integration of the Python library called Xarray allows efficient and easy access to the data, leveraging the familiarity most climate scientists have with it. This approach, combined with the progressive streaming format, not only enhances the findability, shareability and reusability of the data but also facilitates sophisticated analyses and visualizations from commodity hardware like personal cell phones and computers without the need for large computational resources. By collaborating with climate scientists and domain experts from NASA Jet Propulsion Lab and NASA Ames Research Center, we published more than 2 petabytes of climate data via our interactive dashboards for climate scientists and the general public. Ultimately, our solution fosters quicker decision-making, greater collaboration, and innovation in the global climate science community by breaking down barriers imposed by hardware limitations and geographical constraints and allowing access to sophisticated visualization tools via publicly available dashboards.

Managing Large-scale Atmospheric and Oceanic Climate Data for Efficient Analysis and On-the-fly Interactive Visualization

Aashish Panta, Valerio Pascucci, Giorgio Scorzelli, Amy Gooch, Nina McCurdy*, David Ellsworth*, Kyo Lee**

University of Utah, NASA Ames Research Center*, NASA JPL**

PRESENTED AT:

AGU24
Washington, D.C. | 9–13 December 2024

INTRODUCTION

Traditional tools are often unable to handle massive datasets creating bottleneck in data analysis and visualization. We present a groundbreaking visualization dashboard designed to address these issues by enabling progressive streaming and visualization.

1.Scalable Visualization Framework: An innovative, scalable dashboard framework that enable progressive visualization with advanced analytical tools.

2.Efficient Data Reduction and Optimization: A framework that enables efficient storage and transfer of large- scale data with analysis-ready cloud optimized format.

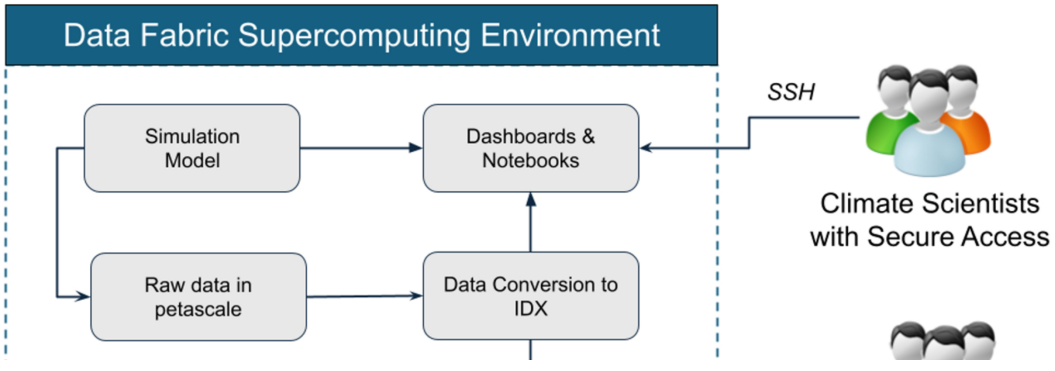
3.Data Democratization: Successfully converted and migrated more than a petabyte of raw data from Pleiades, a NASA supercomputer to the cloud, enhancing public access and collaborative opportunities.

WORKFLOW

1.Our dashboards are dynamically adjusted to fetch the data from the appropriate source as needed.

2.For NASA scientists who have access to Pleiades, a NASA supercomputer, the dashboard will fetch the data from their own filesystem.

3.Other users fetch the data from the cloud which will be cached as needed in order to reduce data transmission over networks.



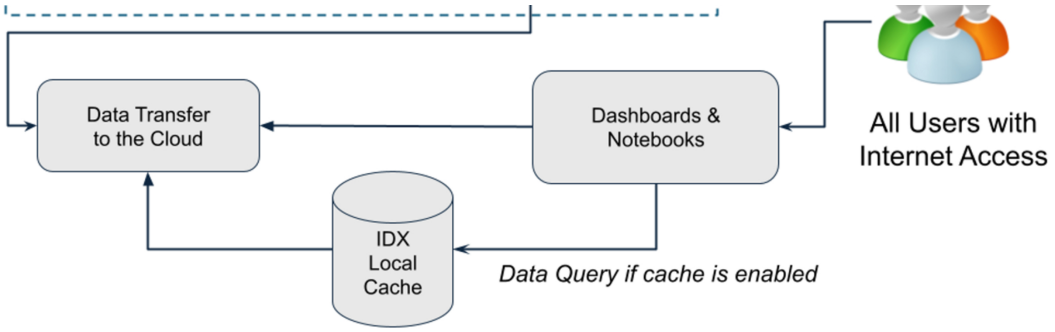


Fig 1: An example workflow showing data conversion and retrieval pipeline for Jupyter notebook and dashboards from a supercomputing environment.

INTERACTIVE VISUALIZATION AND DASHBOARDS

• We worked on two large-scale climate simulation datasets: DYAMOND dataset and LLC4320 Ocean dataset, with several variables like ocean velocity, salinity, sea surface temperature.

• We developed an innovative data conversion pipeline to convert more than a petabyte of this data to OpenVisus IDX format, known for its fast and progressive streaming capabilities. We also tested various compression algorithms, both lossy and lossless, on these datasets

• In an effort to make the data publicly available, we uploaded over 1.2 PB of compressed data to SealStorage, a decentralized cloud storage service. We deployed several dashboards and Jupyter notebooks fetching the cloud data from the cloud storages.

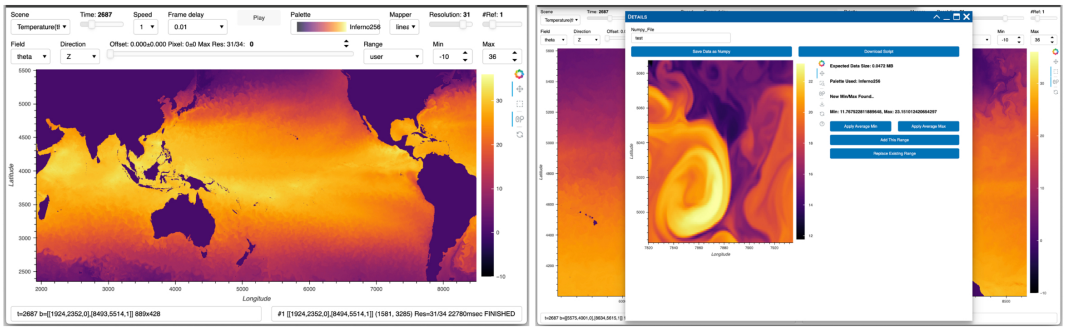


Fig 2: Interactive visualization of sea surface temperature, along with inset for selecting the regions of interest (left). The dashboard provides the ability to directly download the data locally or to download a Python script that fetches the region from the cloud (right).

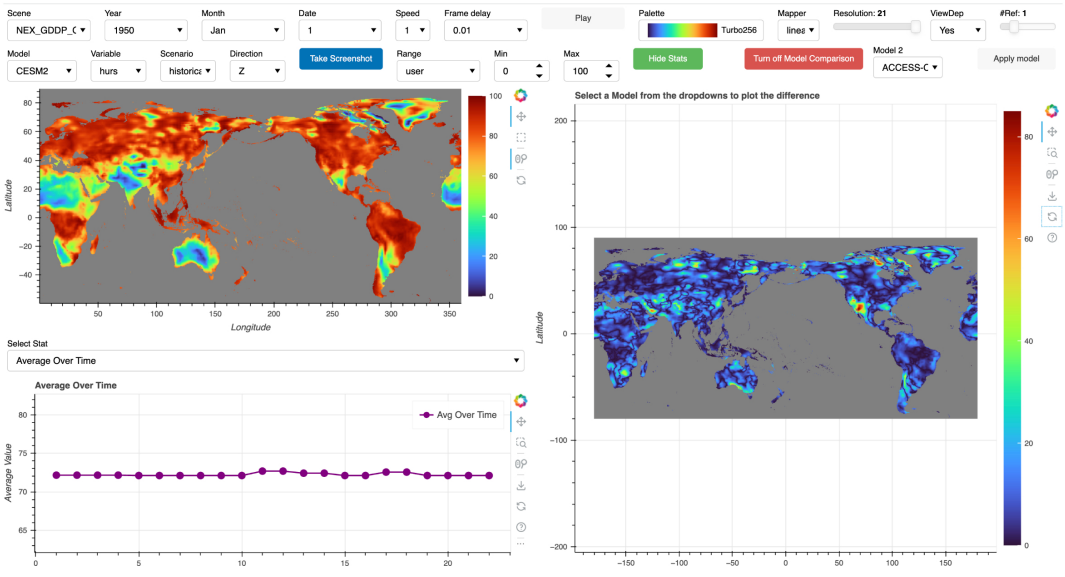


Fig 3: An interactive dashboard showing side by side comparison of multiple models at once. Other general stats like Average Over Time, min/max, are also available.

EXAMPLES

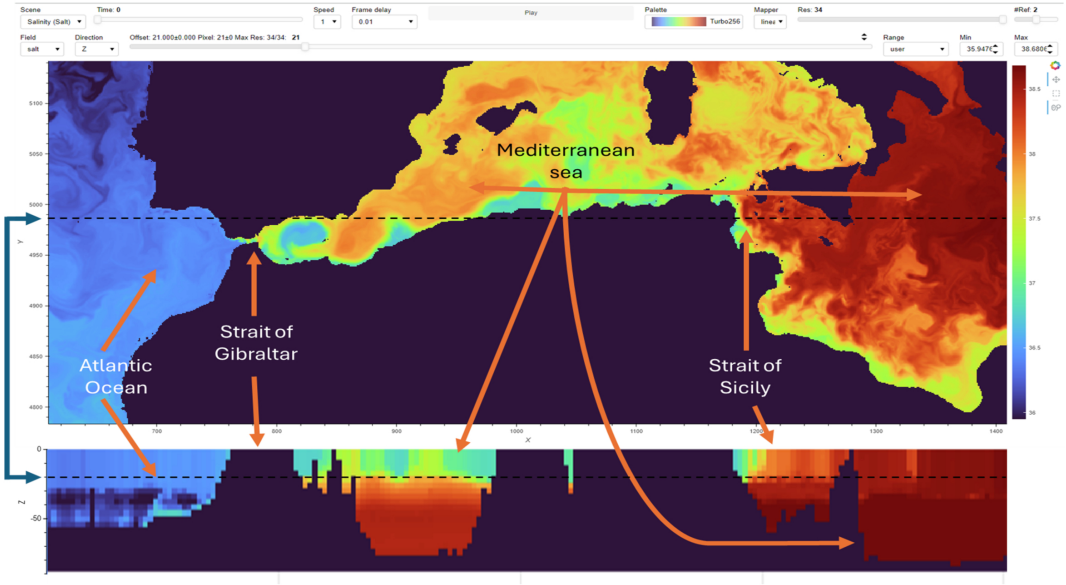


Fig 4: Zoomed-in view of the general water circulation through the Strait of Gibraltar connecting the Mediterranean with the Atlantic Ocean. The low-salinity water enters the Mediterranean Sea from the Atlantic through the Strait of Gibraltar; then the salinity increases, and the water starts to sink as the current moves east. This type of on-the-fly selection of interesting regions from a massive dataset and playing through the time facilitates a deeper understanding of complex climatic phenomena, which was not practically accessible before the implementation of our framework.

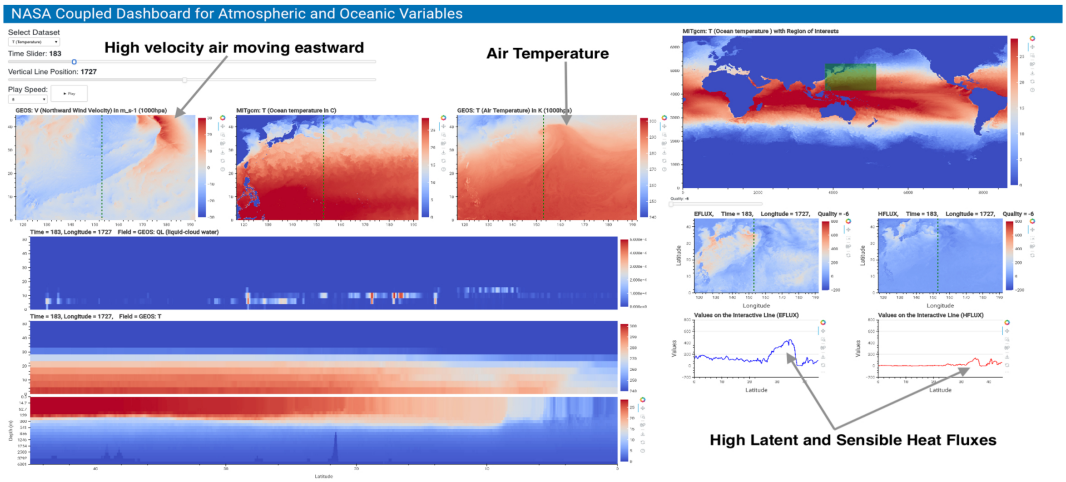


Fig 5:Increasing heat fluxes (two plots and images at the bottom right) and air temperature (image at top middle) creating a high-velocity wind (image at top left) in the atmosphere moving eastward for the Kuroshio region.

PERFORMANCE METRICS

• After converting and moving the data to cloud, we tested the performance of our architecture to calculate the total time

it takes to load the data from NASA Filesystem and our local machine, both cached and not cached.

• We observe that the time it takes to load the data from the user's computer after caching is very close to loading it directly from the Pleiades. Also, the gap continue to decrease as we decrease the resolution (Tested on M1 Macbook Pro 2020, 16GB RAM).

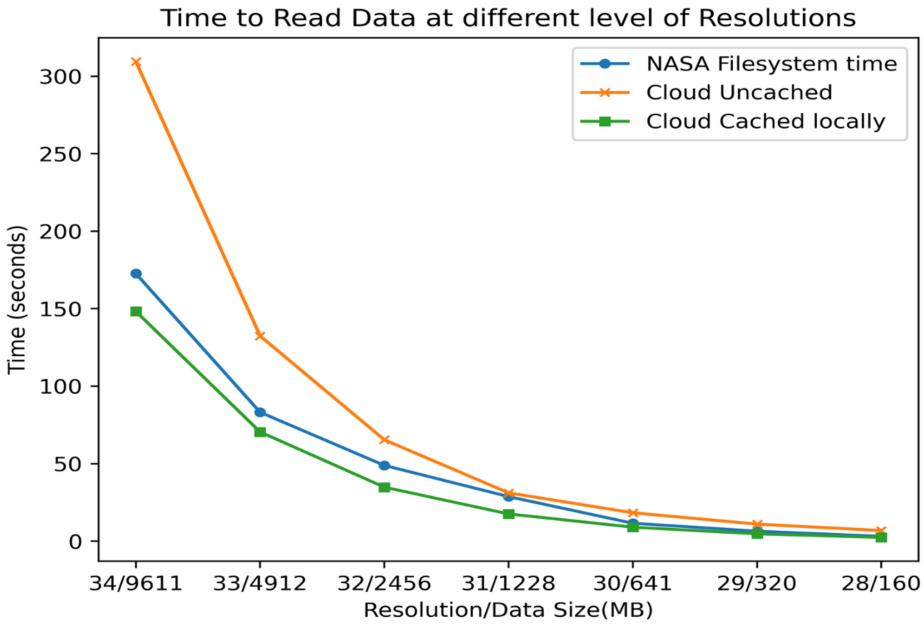


Fig 6: A line plot showing the total time it takes to retrieve a single timestep data from different sources, along with the data size

QR CODE TO DASHBOARD AND CONTACTS

aashish.panta@utah.edu

valerio.pascucci@utah.edu

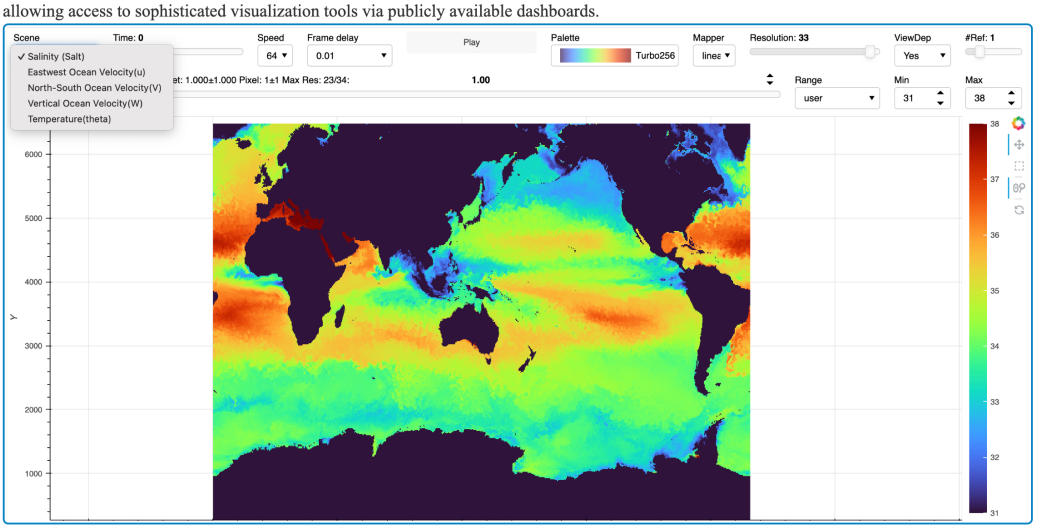
AUTHOR INFO

Aashish Panta is a Ph.D. student at the University of Utah, Salt Lake City, and a graduate research assistant at the Scientific Imaging and Computing Institute. His research focuses on developing visualization tools for large-scale scientific data, with interests in cloud computing and machine learning.

TRANSCRIPT

ABSTRACT

Managing vast volumes of climate data, often reaching into terabytes and petabytes, presents significant challenges in terms of storage, accessibility, efficient analysis, and on-the-fly interactive visualization. Traditional data handling techniques are increasingly inadequate for the massive atmospheric and oceanic data generated by modern climate research. We tackled these challenges by reorganizing the native data layout to optimize access and processing, implementing advanced visualization algorithms like OpenVisus for real-time interactive exploration, and extracting comprehensive metadata for all available fields to improve data discoverability and usability. Our work utilized extensive datasets, including downscaled projections of various climate variables and high-resolution ocean simulations from NEX GDDP CMIP6 and NASA DYAMOND datasets. By transforming the data into progressive, streaming-capable formats and incorporating ARCO (Analysis Ready, Cloud Optimized) features before moving them to the cloud, we ensured that the data is highly accessible and efficient for analysis, while allowing direct access to data subsets in the cloud. The direct integration of the Python library called Xarray allows efficient and easy access to the data, leveraging the familiarity most climate scientists have with it. This approach, combined with the progressive streaming format, not only enhances the findability, shareability and reusability of the data but also facilitates sophisticated analyses and visualizations from commodity hardware like personal cell phones and computers without the need for large computational resources. By collaborating with climate scientists and domain experts from NASA Jet Propulsion Lab and NASA Ames Research Center, we published more than 2 petabytes of climate data via our interactive dashboards for climate scientists and the general public. Ultimately, our solution fosters quicker decision-making, greater collaboration, and innovation in the global climate science community by breaking down barriers imposed by hardware limitations and geographical constraints and allowing access to sophisticated visualization tools via publicly available dashboards.



EVALUATIONS

#	Average Score
There are currently no completed evaluations for this presentation	