



**NCCS**

# **NCCS User Forum**

**30 March 2010**



# Agenda

NCCS

**Welcome & Introduction**  
**Lynn Parnell/Phil Webster**

User Services Updates  
Tyler Simon, User Services

Current System Status  
Fred Reitz, HPC Operations

Scaling Jobs with MPI on Discover  
Bill Putman, SIVO ASTG

NCCS Compute Capabilities  
Dan Duffy, Lead Architect

Analysis Updates & 3D Demo  
Tom Maxwell, Analysis Lead

Questions and Comments  
Lynn Parnell/Phil Webster



# Key Accomplishments

## NCCS

- Installed SCU6 (4096 Nehalem cores, 46 TFLOPs additional)
- Moved SCU5 into S100 computer room and coupled SCU5 and SCU6 I/O fabrics
- Demonstrated use of >4100 cores across SCU5 and SCU6
- Augmented Discover SAN storage
- ARRA-funded Visualization Room



NCCS

# NCCS Staff Additions

- Yingshuo Shen, Earth System Grid (ESG) Data Node
- Laura Carriere, ESG Data Node (part time)



# Agenda

NCCS

Welcome & Introduction  
Lynn Parnell/Phil Webster

User Services Updates  
Tyler Simon, User Services

**Current System Status**  
**Fred Reitz, HPC Operations**

Scaling Jobs with MPI on Discover  
Bill Putman, SIVO ASTG

NCCS Compute Capabilities  
Dan Duffy, Lead Architect

Analysis Updates & 3D Demo  
Tom Maxwell, Analysis Lead

Questions and Comments  
Lynn Parnell/Phil Webster



# Key Accomplishments

## NCCS

### ***Discover/Analysis Environment***

- Upgraded operating system to SLES10 SP2 on all IBM (Harpertown, Nehalem) nodes
- Installed *SCU6* (cluster totals: 14,968 compute CPUs, 155 TF)
- Increased SAN storage
- Increased allowable maximum CPU count for general (1,024), general high (3,072) queues
- Relocated *SCU5* from E100 to S100 (co-located with *SCU6*)
- Replaced two 288-port InfiniBand switches experiencing internal connector problems

### ***Dataportal***

- Upgraded operating system from SLES10 SP2 to SLES11
- Implemented GDS OPeNDAP performance enhancements
- Implemented GPFS-CNFS for improved NFS mount availability

### **DMF**

- Migrated DMF from Irix to Linux
- Added 8 T10KB tape drives
- Upgraded HSM software

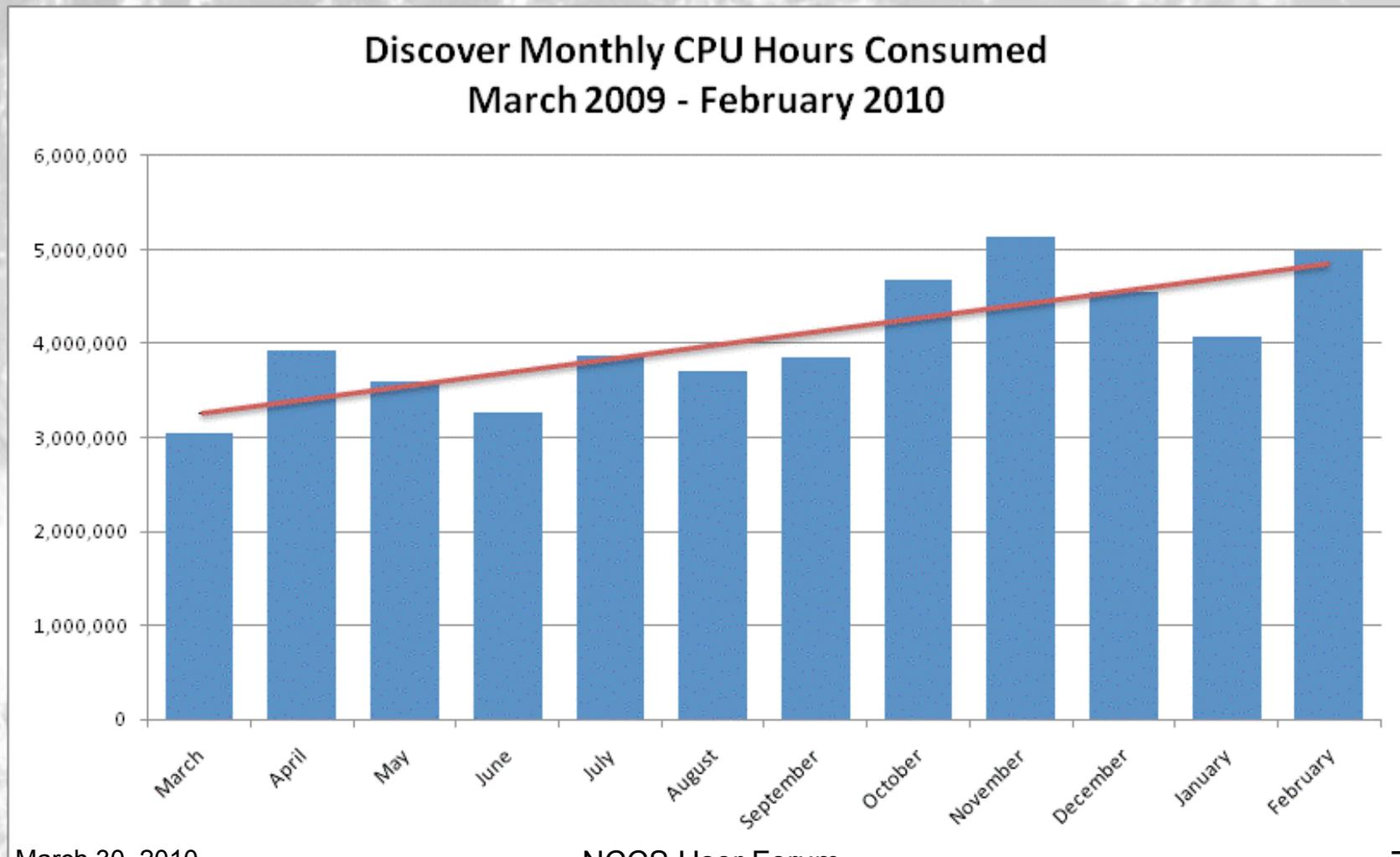
### **Other**

- Completed NCCS physical network reorganization to further enhance network redundancy
- Upgraded NCCS firewall
- Completed *SourceMotel* to *Progress* migration



# Discover Total CPU Consumption Past 12 Months (CPU Hours)

NCCS



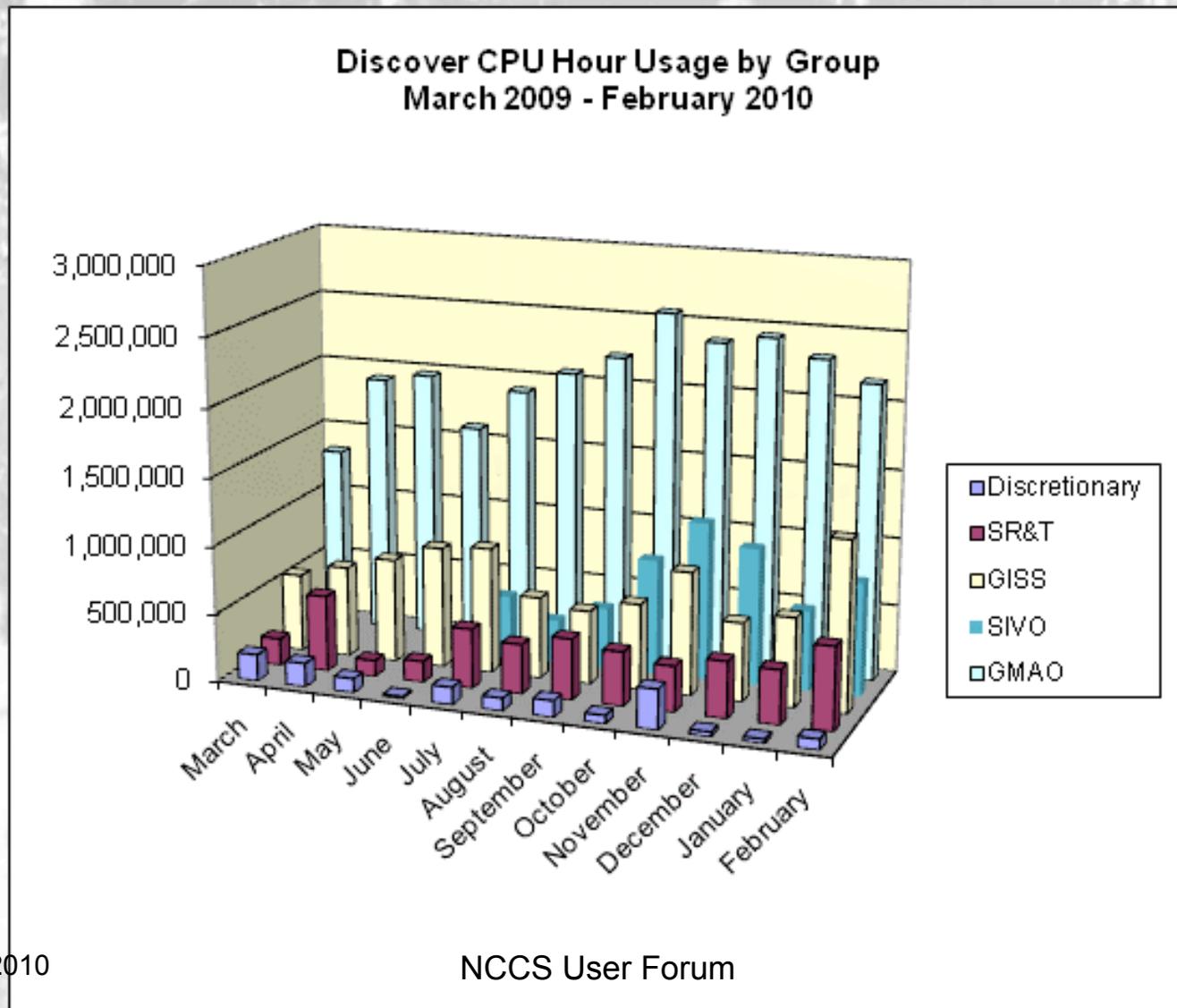
March 30, 2010

NCCS User Forum



# Discover CPU Consumption by Group Past 12 Months (CPU Hours)

NCCS



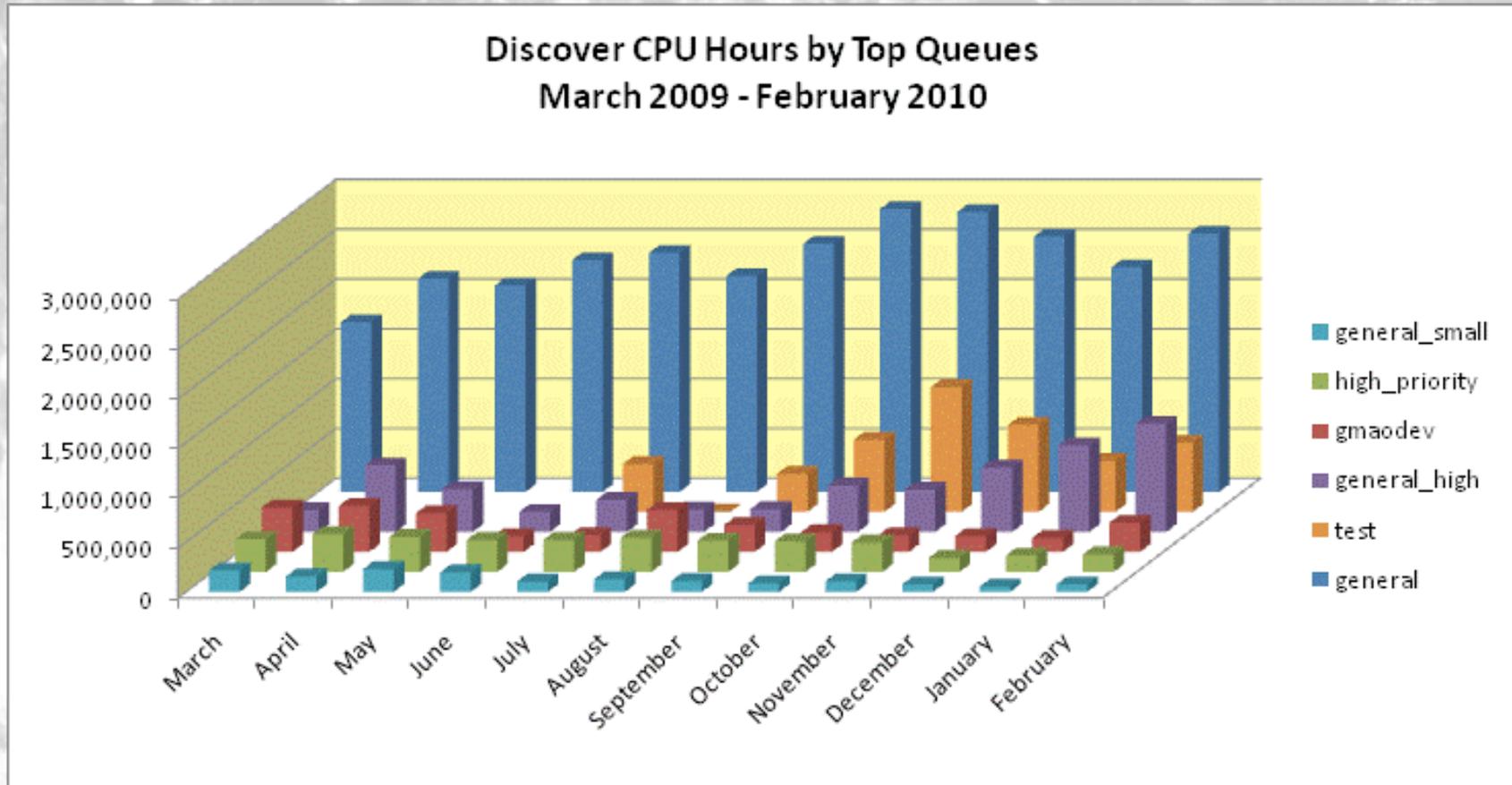
March 30, 2010

NCCS User Forum



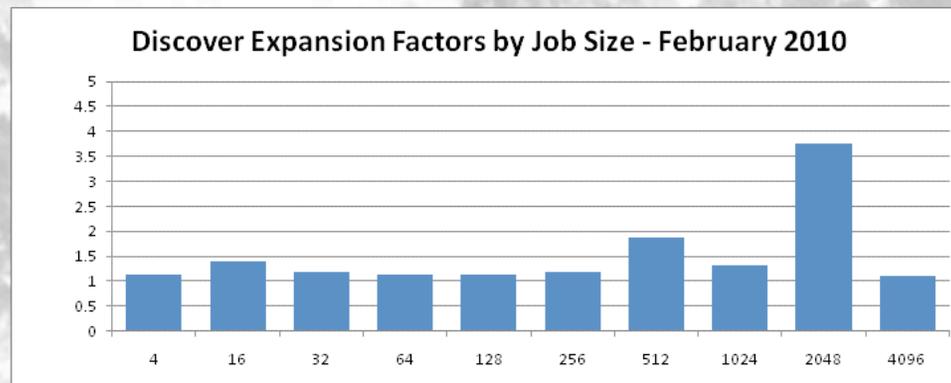
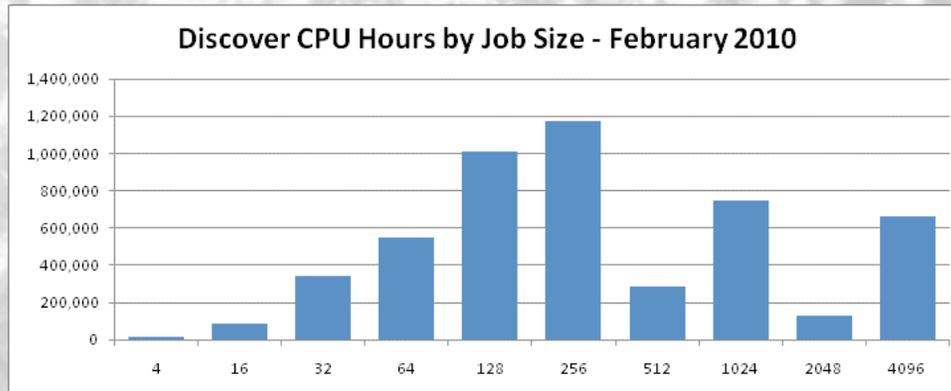
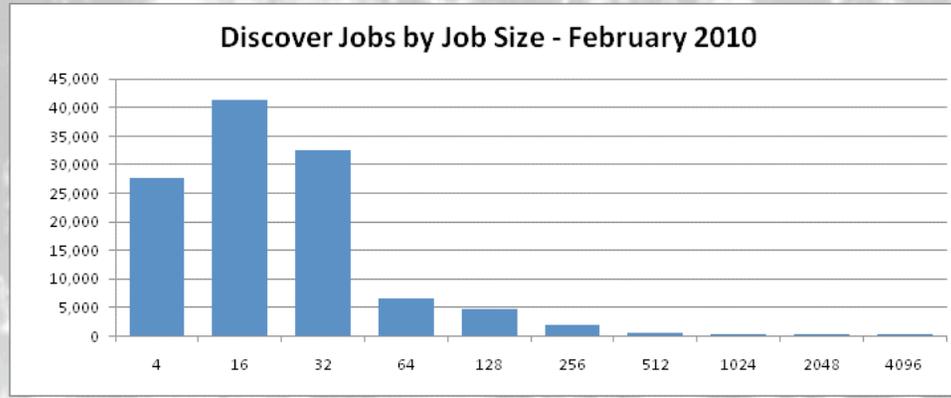
# Discover CPU Consumption by Queue Past 12 Months (CPU Hours)

NCCS



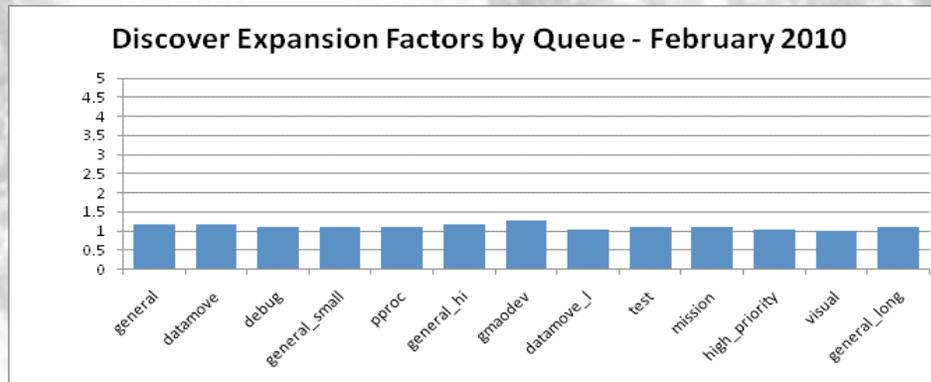
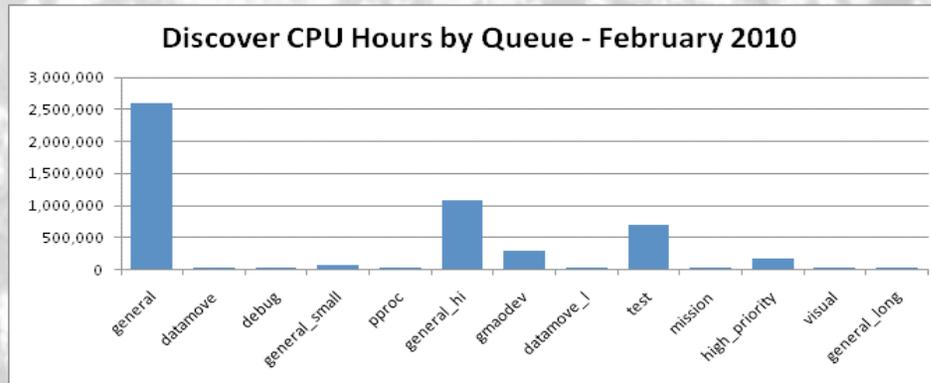
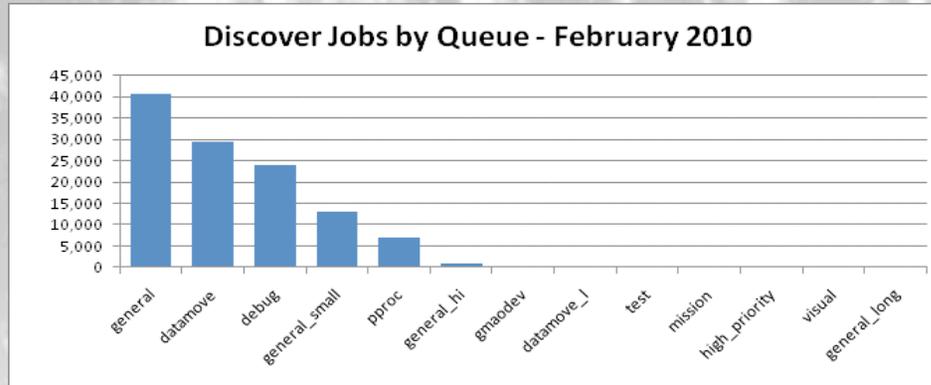


# Discover Job Analysis by Job Size – February 2010





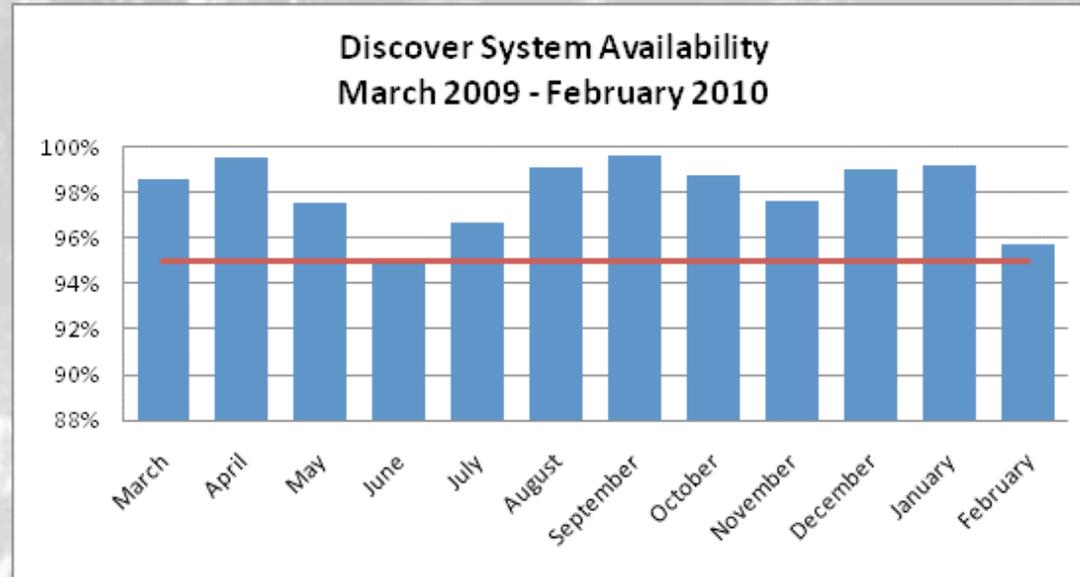
# Discover Job Analysis by Queue – February 2010





NCCS

# Discover Availability – Past 12 Months



## Discover Outages: September – February

### Scheduled Maintenance

- 16 October – 6 hours  
Connect SCU6 to SAN
- 12 November – 14 hours 30 minutes  
SAN disk firmware upgrade
- 16 December – 7 hours 45 minutes  
GPFS 3.2.1-16
- 14 January – 2 hours 20 minutes  
SCU5, SCU6 I/O
- 24 February – 8 hours  
InfiniBand, SAN disk controller replacement

### Unscheduled Outages

- 16 September – 2 hours 54 minutes  
SCU3, SCU4 firmware errors
- 16 October – 3 hours 29 minutes  
Extended maintenance window
- 12 November – 2 hours 39 minutes  
Extended maintenance window
- 9 February – 2 hours 30 minutes  
borgmg server hung
- 11-12 February – 12 hours 15 minutes  
SAN disk controller error
- 24 February – 6 hours 5 minutes  
Extended maintenance window

March 30, 2010

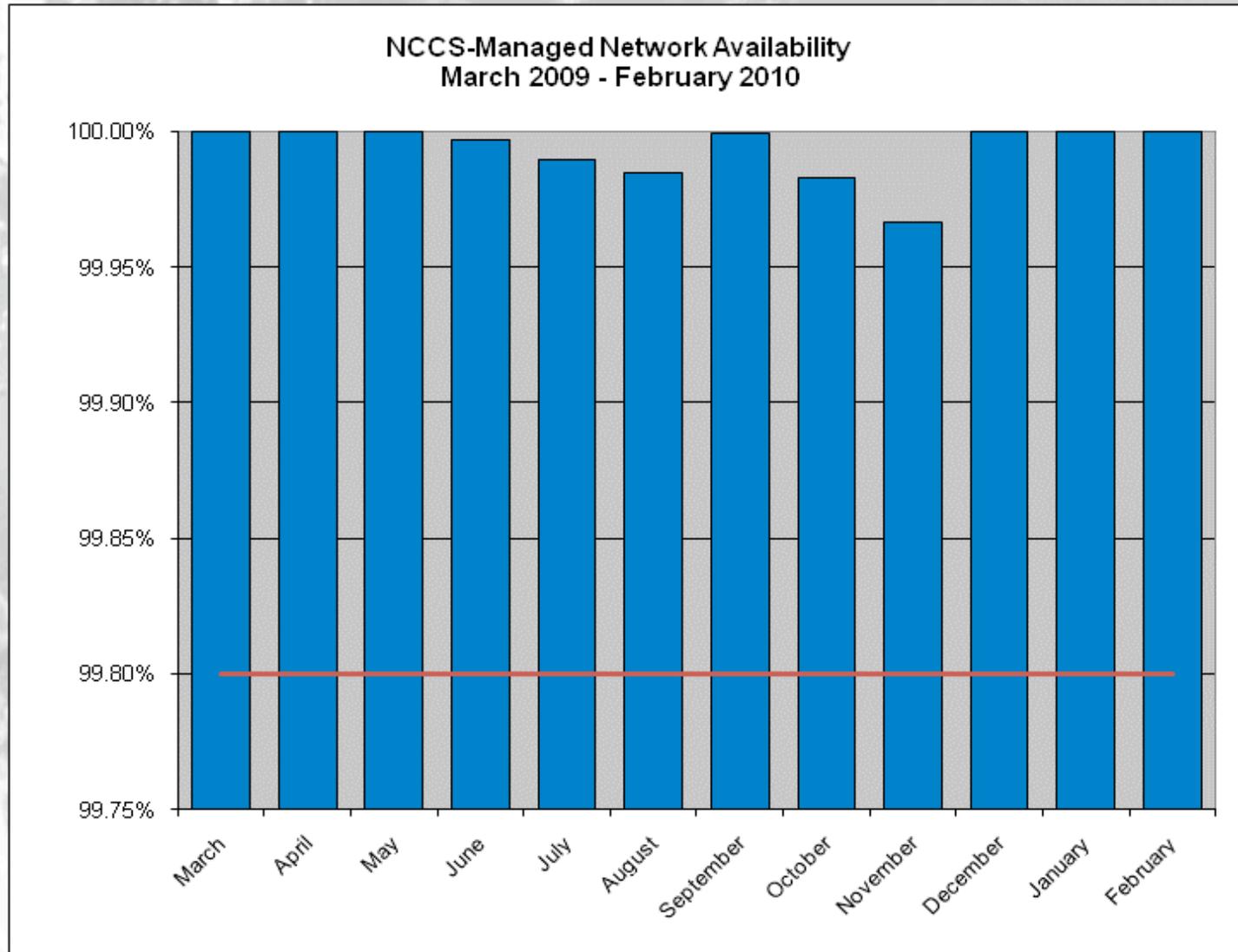
NCCS User Forum

12



NCCS

# NCCS Network Availability – Past 12 Months



March 30, 2010

NCCS User Forum

13



NCCS

## Current *Discover* Issues: Cron Node Hangs/Crashes

- **Symptom:** Cron node becomes unresponsive.
- **Impact:** Cron jobs are no longer launched, users cannot login to cron node.
- **Status:** Contacting users running large work or hundreds of processes via cron to explore other options. Investigating other cron node options or enhancements.



# Future Enhancements *Discover Cluster*

NCCS

- PBS v10
- Reorganize storage
- Increase storage for users and projects (in progress)
- Upgrade I/O nodes serving *Base Unit* and *SCU1-SCU4*
- Implement major software stack upgrade
  - SLES 11
  - OFED 1.5.1
  - GPFS 3.3



# Agenda

NCCS

Welcome & Introduction  
Lynn Parnell/Phil Webster

User Services Updates  
Tyler Simon, User Services

Current System Status  
Fred Reitz, HPC Operations

Scaling Jobs with MPI on Discover  
Bill Putman, SIVO ASTG

**NCCS Compute Capabilities**  
**Dan Duffy, Lead Architect**

Analysis Updates & 3D Demo  
Tom Maxwell, Analysis Lead

Questions and Comments  
Lynn Parnell/Phil Webster



NCCS

# SCU5 & SCU6 and Other Discover Augmentations

- SCU5 and SCU6 were integrated into the same room during February and March
  - Single Infiniband (IB) fabric capable of running jobs up to 8K cores
  - Users who want to scale large jobs should contact User Services
- Demonstrated scaling up to 4K cores by Bill and Tyler
  - MVAPICH2 installed to help with scaling
  - This will be made into a generally available module soon
- All 8K cores will be made available to the general queue by April 1<sup>st</sup>
- Significant disk capacity upgrades were implemented
  - Large amount have been allocated to such projects as the IPCC runs
  - Additional capacity is available; please contact User Services



# FY10 Compute Upgrade

## NCCS

- Expand compute nodes within the SCU5 and SCU6 Infiniband (IB) fabric
  - Compute nodes upgrade only with the same configuration as existing Nehalem nodes with the exception of the type of processor (see below)
- Slightly different blocking factor within the IB fabric
  - Should not significantly affect application performance
- Expecting delivery of equipment in July 2010
- Initial performance of the GEOS Cubed-Sphere benchmark shows a slight improvement on the new processors
  - Assumes that 4-cores per socket are used on both the Nehalem and the Westmere
  - Moving to the Westmere to provide more cores at the same cost

Processor	Speed	Cores/Socket	Cache Size
Nehalem	2.8 GHz	4	12 MB
Westmere	2.8 GHz	6	12 MB



# NCCS Visualization Room

## NCCS

- 3x5 Hyperwall
  - Dell servers with Nvidia FX1700 GPUs
  - Samsung 46-inch, very small bezel televisions
  - Linux software from the SVS is capable of displaying images across the Hyperwall
  - Also have Windows software that we are exploring for use
- Scientific Workstations
  - Dell workstations with Nvidia FX4800 GPUs
  - Second monitor is a high-definition Samsung monitor attached to the Nvidia GPU
  - 3D capable (or it will be!)
- These capabilities are for all NCCS users
  - Still integrating the systems; fully available by June 1, 2010
  - Please contact User Services for more information



NCCS

# NCCS Visualization Room Visual



March 3

20



# Dataportal Upgrades

NCCS

- Additional disk capacity for the Dataportal
  - 90 TB
  - Integrated into the GPFS environment on the Dataportal
  - To be available by Summer 2010
- Database servers and database storage
  - The NCCS is creating a database service for the Dataportal
  - Additional Dell servers and storage set aside for a Postgres database will be added to the Dataportal
  - To be available by Summer 2010
- Contact User Services if you are in need of any general data distribution capabilities, need more capacity on the portal, or desire more information about the database service.



# Archive Upgrades (Fall 2010)

NCCS

- Additional storage
  - Increased disk cache: Larger file systems using significantly faster disk
- Fewer, larger file systems
  - Allows more files to remain on disk for longer periods
  - Decreased load and tape drive wait times
- Enhanced NFS support
  - SGI's enhanced NFS: Addresses issues with standard SLES NFS server
- Upgraded DMF servers
  - Still working out the details
- Software stack upgrade
  - Newer version of DMF with SLES11

March 30, 2010

NCCS User Forum

22



# Agenda

NCCS

Welcome & Introduction  
Lynn Parnell/Phil Webster

**User Services Updates**  
**Tyler Simon, User Services**

Current System Status  
Fred Reitz, HPC Operations

Scaling Jobs with MPI on Discover  
Bill Putman, SIVO ASTG

NCCS Compute Capabilities  
Dan Duffy, Lead Architect

Analysis Updates & 3D Demo  
Tom Maxwell, Analysis Lead

Questions and Comments  
Lynn Parnell/Phil Webster



**NCCS**

# User Services

The NCCS began its Fiscal Year 2010 (on November 1, 2009) with 546 valid users. Since that time...

25 users have been removed from NCCS access:

- 14 on the discover cluster,
- 4 on the Data Analysis (dali) nodes on the discover cluster and
- 15 on the Data Management Facility on dirac.

51 users have been added to NCCS User Community:

- 45 on the discover cluster,
- 45 on the Data Analysis (dali) nodes on the discover cluster and
- 45 on the Data Management Facility on dirac.

At the present there are 582 valid NCCS users:

- 502 on the discover cluster,
- 289 on the Data Analysis (dali) nodes on the discover cluster and
- 526 on the Data Management Facility on dirac.



NCCS

# Ticketing System Upgrade

- Access the NCCS ticketing system via a web interface.
- Create and edit your own tickets and search the built-in knowledge database.
- Attach files directly to your tickets via web.



NCCS

# Website Updates

- Batch Queue Updates

[http://www.nccs.nasa.gov/discover\\_queues.html](http://www.nccs.nasa.gov/discover_queues.html)

- Intel MPI programming and optimization guides

[http://www.nccs.nasa.gov/discover\\_qna.html#step32](http://www.nccs.nasa.gov/discover_qna.html#step32)

- Discover Job Monitor

<http://www.nccs.nasa.gov/jobmon>

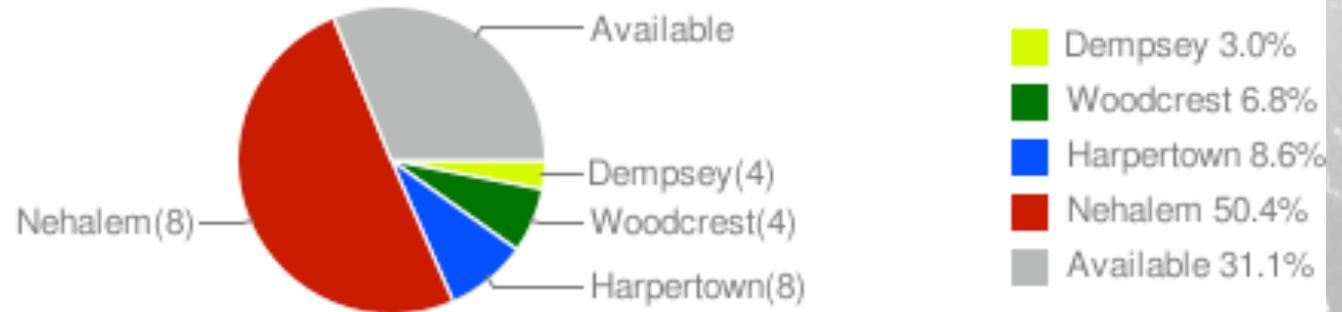
- New Website Look & Feel...coming soon



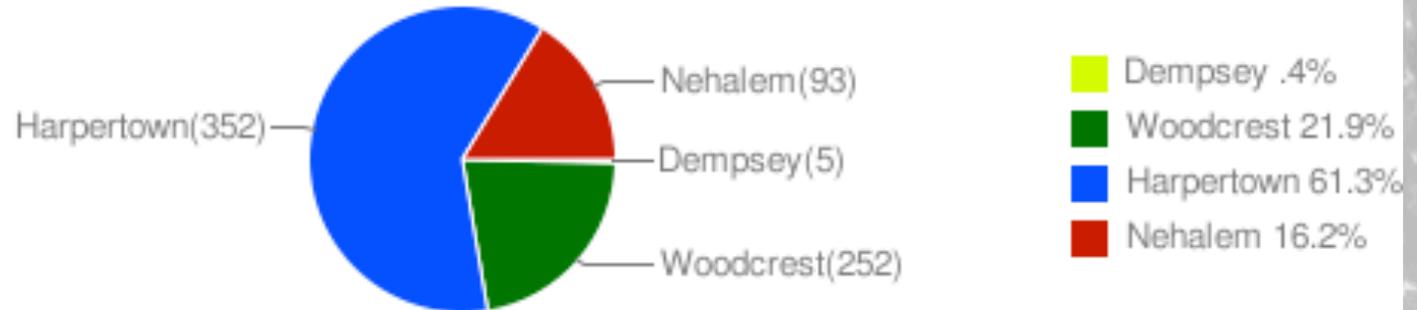
NCCS

# Discover Job Monitor Updates

System Utilization by Node Type (cores per node)



Available Nodes by Type (nodes free)



Job Wait times in quantiles from previous 2 days

March 30, 2010

NCCS User Forum

27



# Agenda

NCCS

Welcome & Introduction  
Lynn Parnell/Phil Webster

User Services Updates  
Tyler Simon, User Services

Current System Status  
Fred Reitz, HPC Operations

**Scaling Jobs with MPI on  
Discover  
Bill Putman, SIVO ASTG**

NCCS Compute Capabilities  
Dan Duffy, Lead Architect

Analysis Updates & 3D Demo  
Tom Maxwell, Analysis Lead

Questions and Comments  
Lynn Parnell/Phil Webster



# Scaling jobs with MPI on Discover

NCCS

- IntelMPI
  - Good for typical MPI jobs < 2000 cores
  - Uses numerous environment variables
    - More user control
    - Increased complexity
  - Generally does not scale well beyond 2000 cores
- MVAPICH
  - Simpler to use
  - Startup issues at 2000+ cores
- MVAPICH2
  - Improved startup mechanism
  - Scales well to 4000+ cores



NATIONAL AERONAUTICS  
AND SPACE ADMINISTRATION

+ NASA Portal  
+ Request an Account  
+ Modeling Guru Home

**Modeling Guru**  
Beta

<https://modelingguru.nasa.gov>



# Modeling Guru Beta

## Sample Script for Running

The following is a sample PBS script that will demonstrate how to set

```
#!/bin/csh -f
# -----
#PBS -l select=4:ncpus=8:proc=harp
#PBS -l walltime=8:00:00
#PBS -S /bin/csh
#PBS -j eo
#PBS -q SomeScript
#PBS -W group_list=SOMEGROUP
# -----

source /usr/share/modules/init/csh
module purge
module load comp/intel-10.1.021 mpi/impi-3.2.011

# updated to use the v2 DAPL devices
# use only one of the below:
# for Connectx IB in the IBM/Harpertown nodes
setenv I_MPI_DEVICE rdssm:ofa-v2-mlx4_0-1
# for older Infiniserv 3 IB interface in woodcrest/dempsey nodes
# setenv I_MPI_DEVICE rdssm:ofa-v2-mthca0-1

#If you run into any problems always run with debugging set to 9 or higher:
setenv I_MPI_DEBUG 9

#Some DAPL specific settings that may help
setenv DAPL_ACK_RETRY 7
setenv DAPL_ACK_TIMER 20
setenv DAPL_RNR_RETRY 7
setenv DAPL_RNR_TIMER 28

# enabled by default, doesn't hurt to specify
setenv I_MPI_FALLBACK_DEVICE disable

#Enable internal optimizations for large jobs
setenv I_MPI_RDMA_SCALABLE_PROGRESS 1

#Increase timeout for startup command
setenv I_MPI_JOB_STARTUP_TIMEOUT 10000

mpirun -perhost 8 -np 32 ./myprogram.x
```

March 30, 2010

## Using IntelMPI on Discover

VERSION 5

Created on: Mar 13, 2009 11:48 AM by [Rahman Syed](#) - Last Modified: Jul 31, 2009 12:42 PM by [Rahman Syed](#)

IntelMPI is one of the various MPI libraries offered by the NCCS on its Discover system. Traditionally, users have developed their applications with the ScalimPI library; however with the newest software and hardware on Discover, IntelMPI can be a better option for many users.

### Setup

First, ensure that you have passwordless SSH setup properly on Discover. Please see this document for more information: [Password-less logins with use of ssh-keygen](#)

Next, load a compiler and IntelMPI module in your environment with the following command:

```
module load comp/intel-10.1.021 mpi/impi-3.2.011
```

### Build

To compile and link your application using the IntelMPI library, simply use the following commands:

```
mpiifort (for Fortran MPI codes)
mpicc (for C MPI codes)
mpicpc (for C++ MPI codes)
```

These scripts will automatically add include directories for compilation purposes, and library directories for linking purposes.

### Run

To run your application with IntelMPI, first acquire a PBS session. With IntelMPI, all of Discover is available for use, whereas with ScalimPI only Dempsey/Woodcrest nodes were available. With IntelMPI, you can choose the same nodes ScalimPI supports (with the "scali=true" option in your PBS request) or you can use the Harpertown nodes (with the "proc=harp" option in your PBS request).

Once you've acquired a PBS session, issue the following command:

```
mpirun -perhost <cpuspernode> -np <numcpus> ./myprogram.x
```

where is the total number of CPUs you'd like to run with.



## Modeling Guru Beta

[NASA Modeling Guru](#) > [Languages, Libraries & Tools](#) > Documents

[^ Up to Documents in Languages, Libraries & Tools](#)



### Running Applications with MVAPICH

VERSION 1

Created on: Mar 23, 2010 10:09 AM by [Jules Kouatchou](#) - Last Modified: Mar 23, 2010 10:16 AM by [Jules Kouatchou](#)

If you compile your application with MVAPICH, here is the setting you need to introduce in your PBS script:

```
#PBS -l select=4:ncpus=8:mpiprocs=8  
module purge  
module load allYourModules  
mpirun_rsh -np <number of processes> -hostfile $PBS_NODEFILE <executable>
```

Note that in the above "mpiprocs=8" was introduced because with MVAPICH the nodefile has one line per process rather than one host per line like Intel/MPI. It was assume that the application will run on 4 nodes ("select=4"). Any number of nodes will be fine.



# Modeling Guru Beta

## Using MVAPICH2 on Discover

VERSION 1

Created on: Mar 26, 2010 10:26 AM by [wputman](#) - Last Modified: Mar 26, 2010 10:37 AM by [wputman](#)

Many MVAPICH 2 Builds available for various Intel compilers, use `module avail` to view them:

```
other/mpi/mvapich2-1.4.1/intel-9.1.042
other/mpi/mvapich2-1.4.1/intel-9.1.052
other/mpi/mvapich2-1.4.1/intel-10.1.017
other/mpi/mvapich2-1.4.1/intel-10.1.018
other/mpi/mvapich2-1.4.1/intel-10.1.023
other/mpi/mvapich2-1.4.1/intel-11.0.083
other/mpi/mvapich2-1.4.1/intel-11.0.083_debug
other/mpi/mvapich2-1.4.1/intel-11.1.038
other/mpi/mvapich2-1.4.1/intel-11.1.056
other/mpi/mvapich2-1.4.1/intel-11.1.056_debug
```

Use `mpiicc`, `mpiifort`, or `mpiicpc` for compilation

PBS options:

```
#PBS -l select=4:ncpus=8:mpiprocs=8
```

Environment variables to use fast/scalable startup mechanism:

```
setenv MV2_FASTSSH_THRESHOLD 1
setenv MV2_NPROCS_THRESHOLD 1
```

MPI run command:

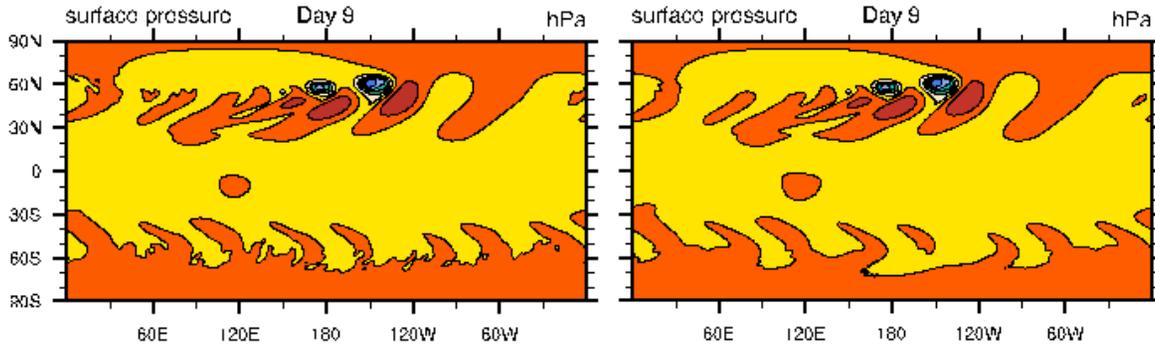
```
mpirun_rsh -np <number of processes> -hostfile $PBS_NODEFILE <executable>
```

Tags: [mpi](#), [mvapich](#), [mvapich2](#), [mpirun\\_rsh](#), [mpiprocs](#)

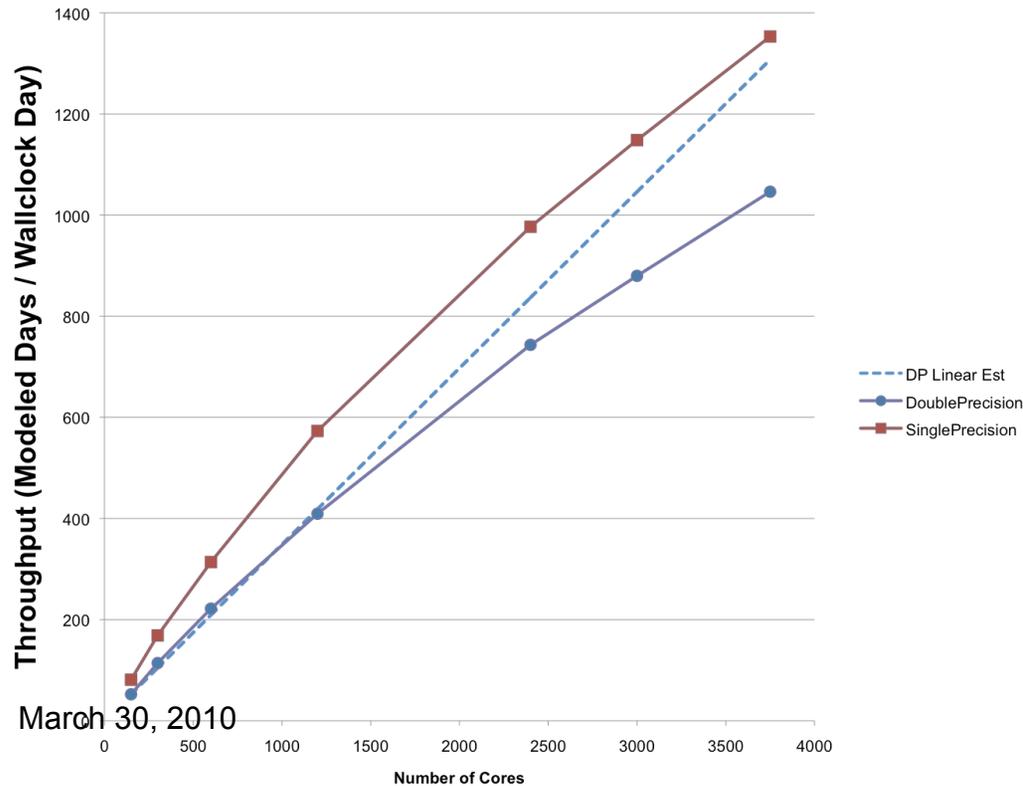
March 30, 2010



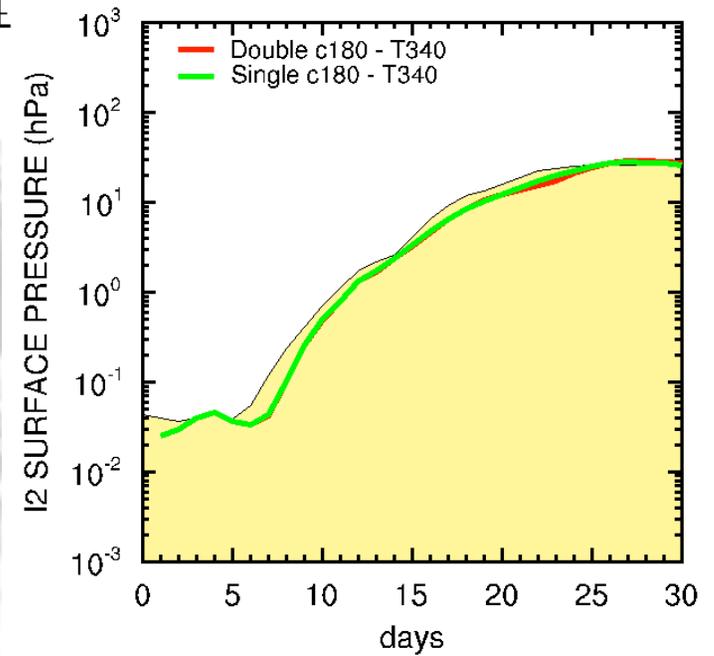
# FV Cubed Dycore (Single Precision?)



20km (c500) 26-Level Cubed-Sphere FV Dycore Benchmark



FVcubed vs EUL SPECTRAL

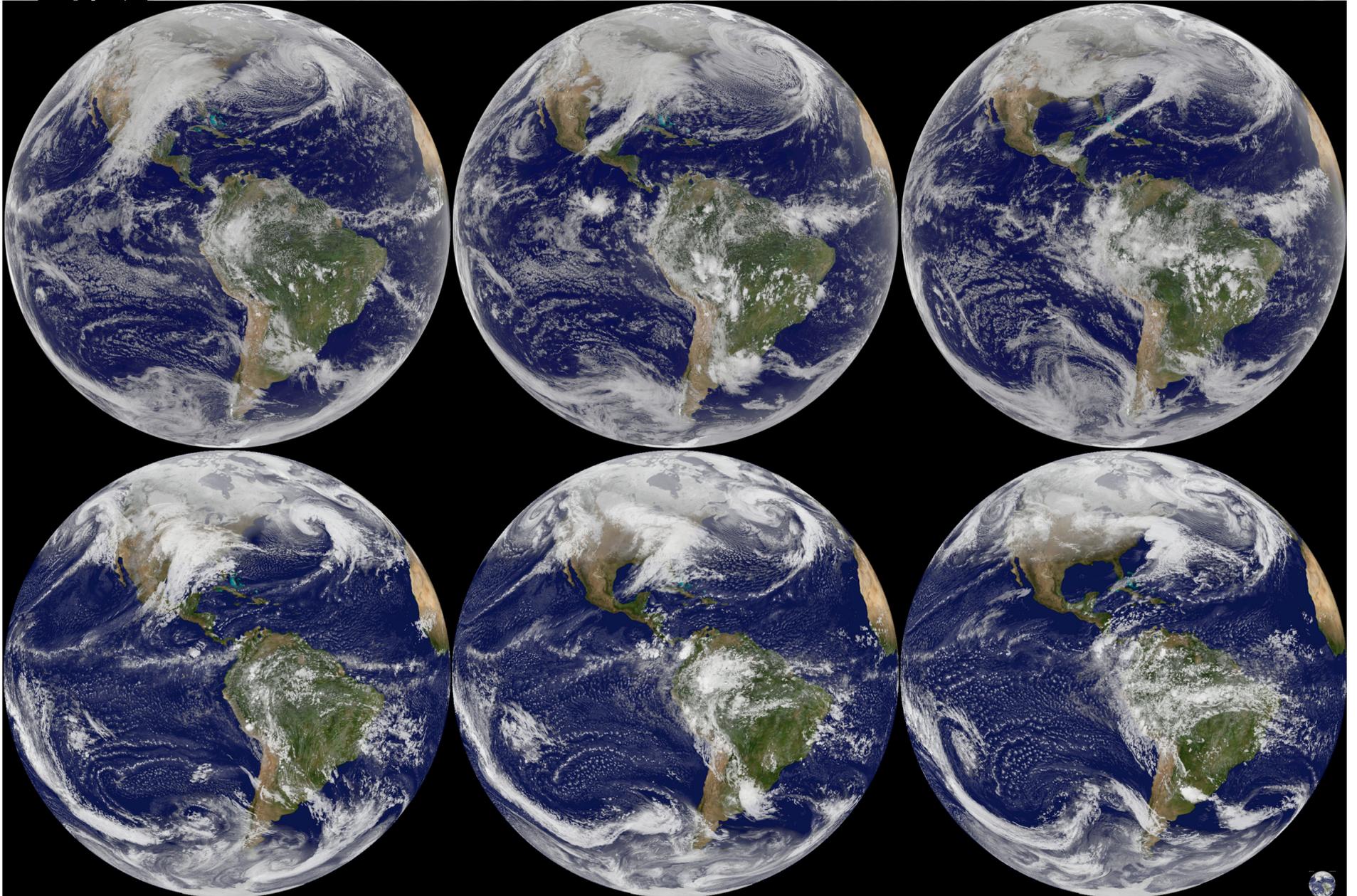


2x - 1.25x  
Speedup



Which is GOES?

Which is GEOS-5?





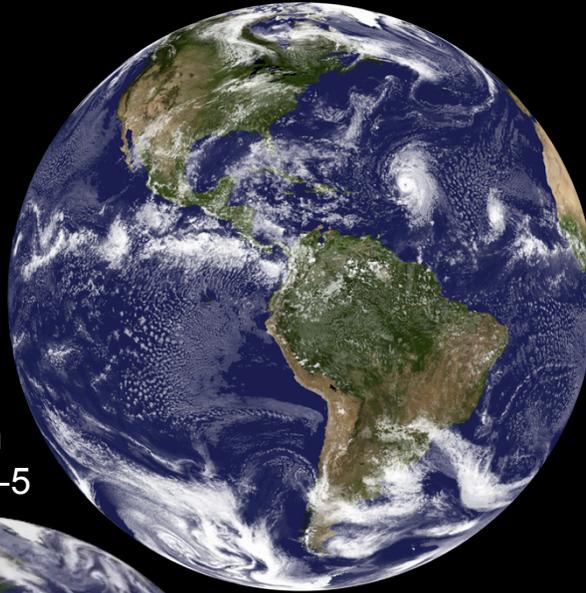
# Hurricane Bill

August 200

GOES



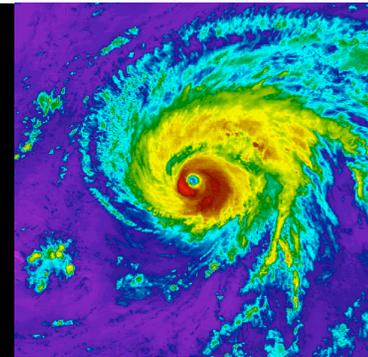
3.5-km GEOS-5



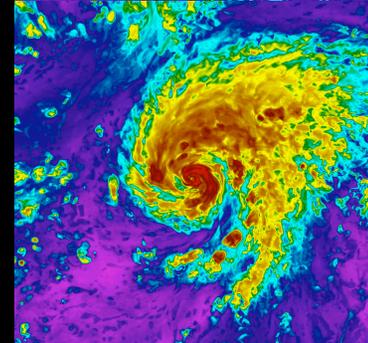
7-km  
GEOS-5



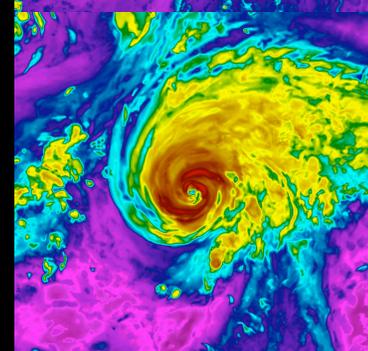
GOES IR



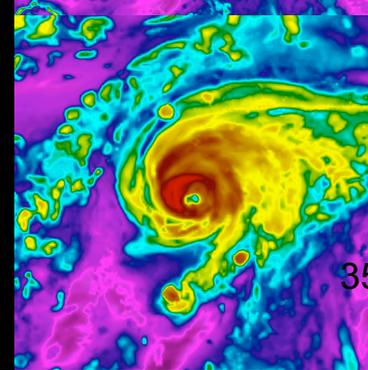
3.5-km GEOS-5



7-km GEOS-5



14-km GEOS-5



4-day Forecast OLR





# Agenda

NCCS

Welcome & Introduction  
Lynn Parnell/Phil Webster

User Services Updates  
Tyler Simon, User Services

Current System Status  
Fred Reitz, HPC Operations

Scaling Jobs with MPI on Discover  
Bill Putman, SIVO ASTG

NCCS Compute Capabilities  
Dan Duffy, Lead Architect

Analysis Updates & 3D Demo

Tom Maxwell, Analysis Lead

Questions and Comments  
Lynn Parnell/Phil Webster



# Earth System Grid Data Node Update

NCCS

- ESG Data Node software (beta) received from PCMDI.
  - Installed and tested on the NCCS Dataportal.
  - Successfully published to PCMDI ESG Gateway.
- ESG (version 1.0) Data Node software due April 1.
- Hired Yingshuo Shen to manage ESG node.
- Testing with YOTC data.
  - Evaluating feasibility of publishing MERRA data.
- Will publish GISS & GMAO IPCC AR5 contributions.
  - Scheduled date for publication of initial data: August 2010



# Analysis Products Server (Under Development)

- Server-side Analysis Services executed on Dali
  - Access via Dataportal
  - Python client for GrADS, CDAT, DV3D, etc.
  - Service APIs:
    - OPeNDAP DAP with extensions
    - Live Access Server (LAS) request protocol
    - IRODS
  - Planned Services
    - Subsetting, regridding/rescaling, reformatting, aggregation, etc.
    - Differencing, averaging (e.g. zonal means)
    - Data transforms ( wavelet, FFT, etc. )



NCCS

# NCAR VAPOR

- Visual data discovery environment
  - tailored for CFD applications
- Desktop application
  - full featured (but complex) GUI
  - capable of handling terascale size data sets
  - wavelet-based multiresolution data representation
  - task-parallel wavelet transform scripts on dali
- Advanced interactive 3D visualization
  - tightly coupled with quantitative data analysis
- Close coupling with IDL



# Vapor User Interface

VAPoR User Interface

File Edit Data View Script Animation Help

Visualizer No. 0

Animation Viewpoint Region Probe Flow DVR

Selected Point Variable at Selected Point

0.319408	0.0861631
0.348329	
0.25	

Attach Selected Point to Flow Seed    Add Selected Point to Flow Seeds

Copy selected point to:

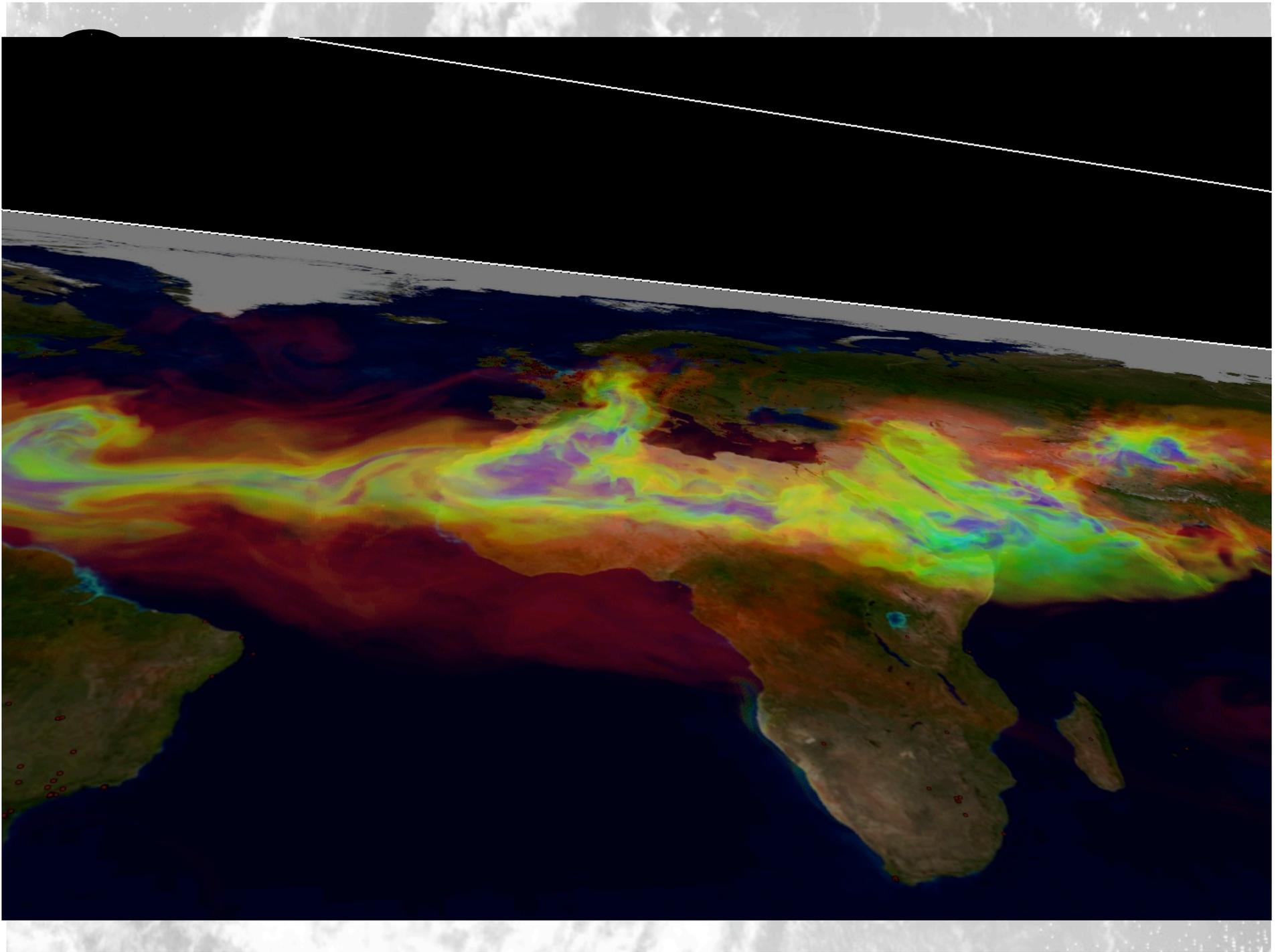
Region Center    View Center    Rake Center    Probe Center

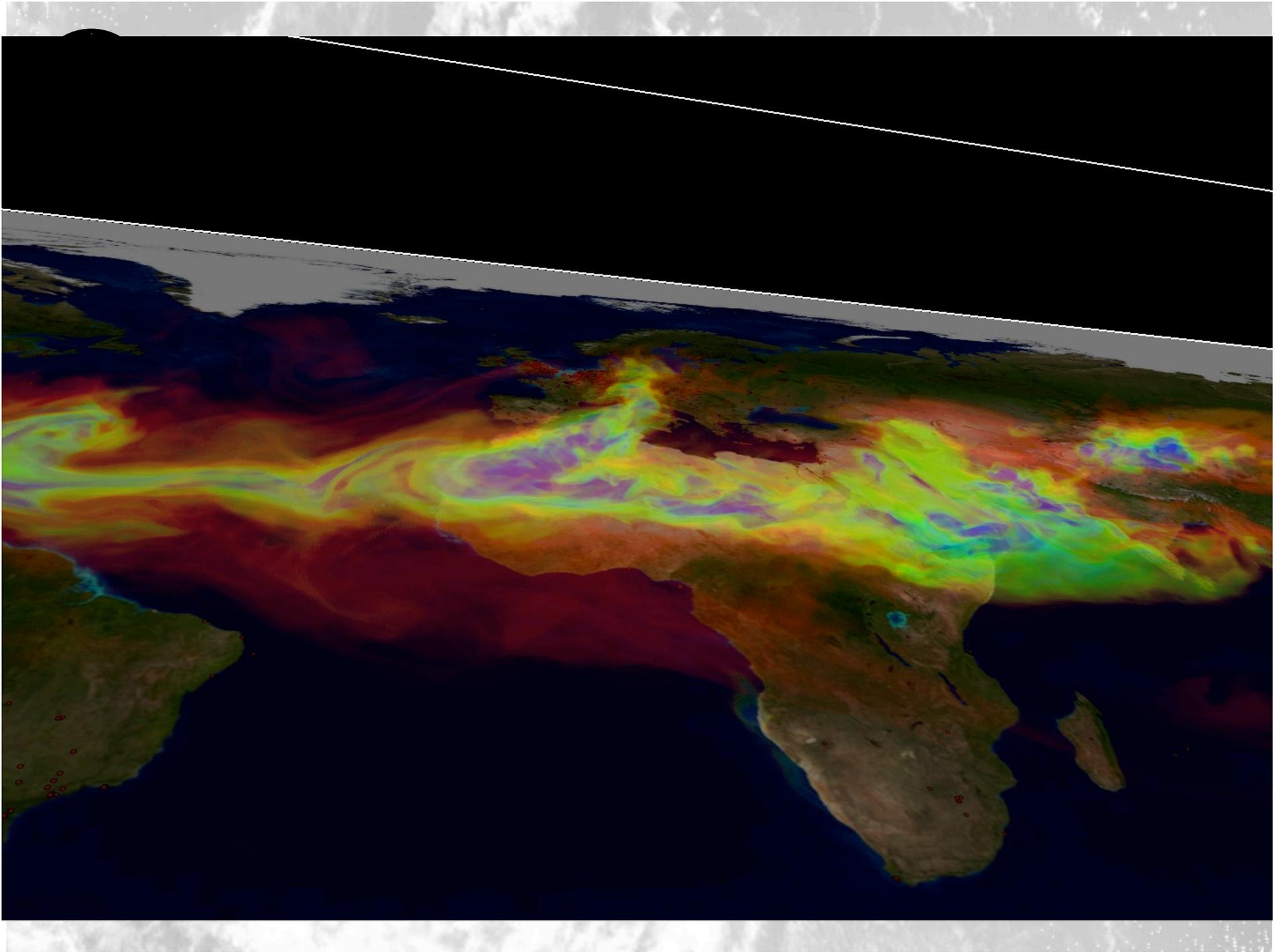
Transfer Function Editor

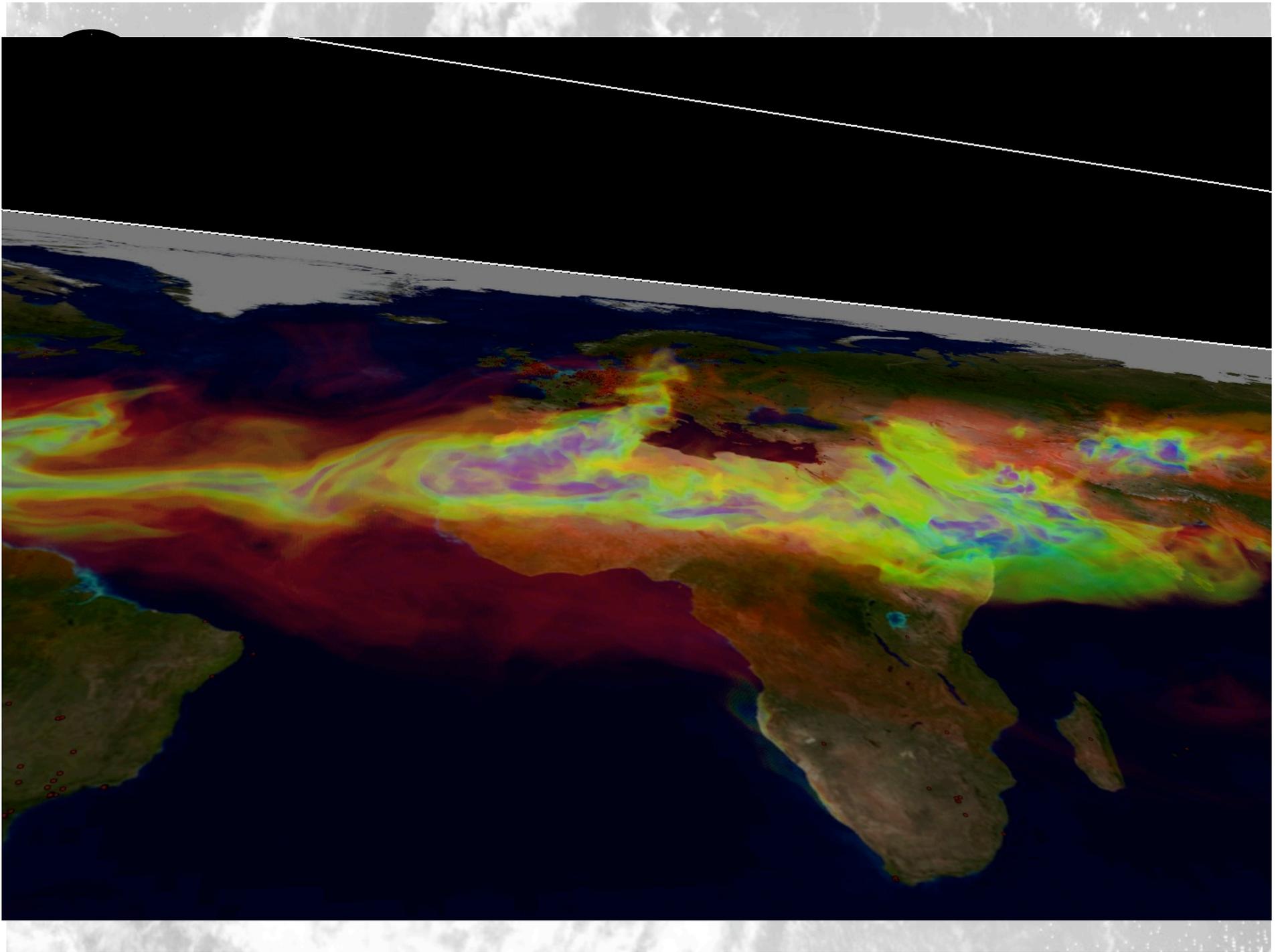
Edit    Zoom/Pan    Fit to View    Histo

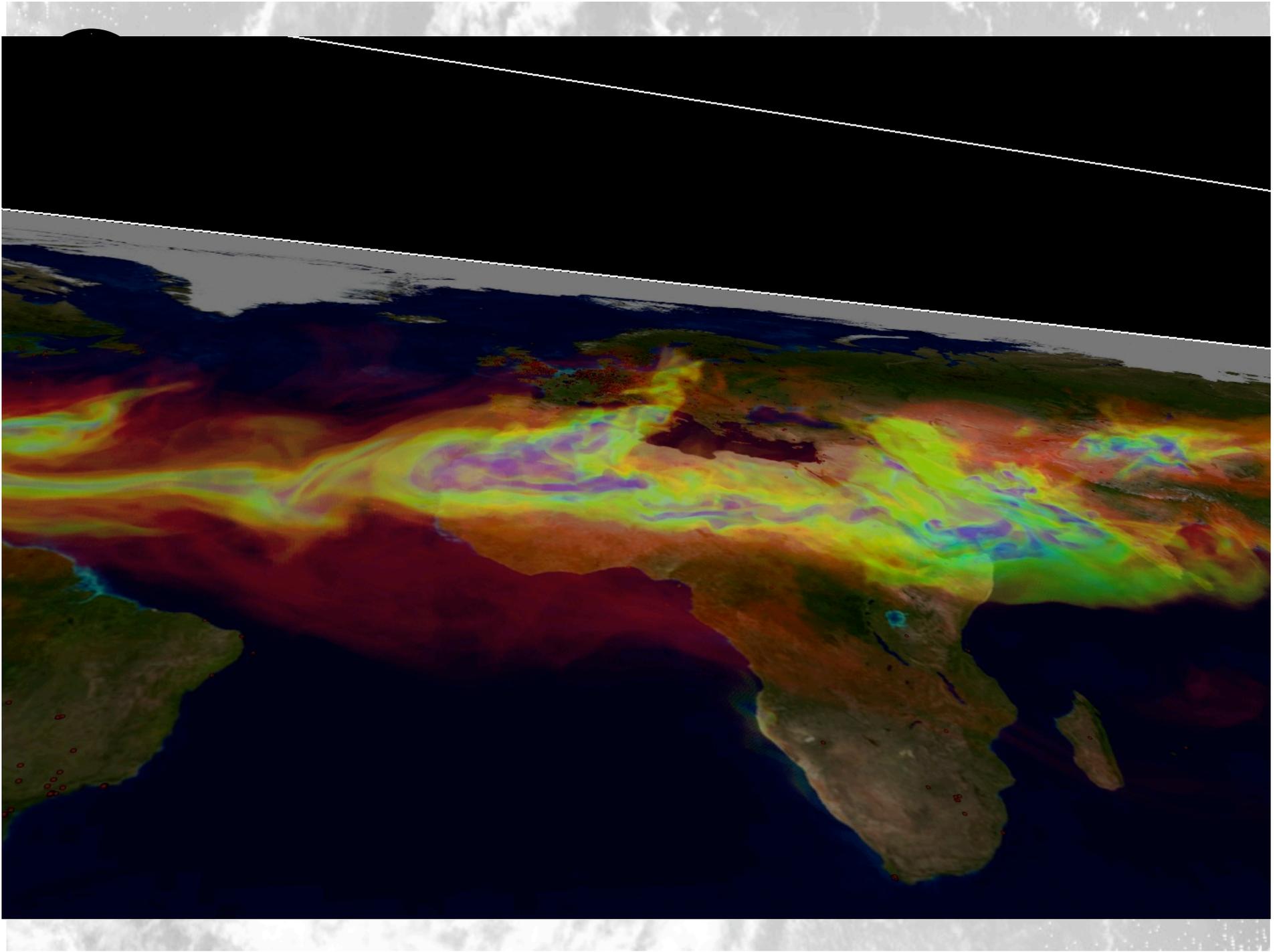
Visualizer No. 0    Visualizer No. 1

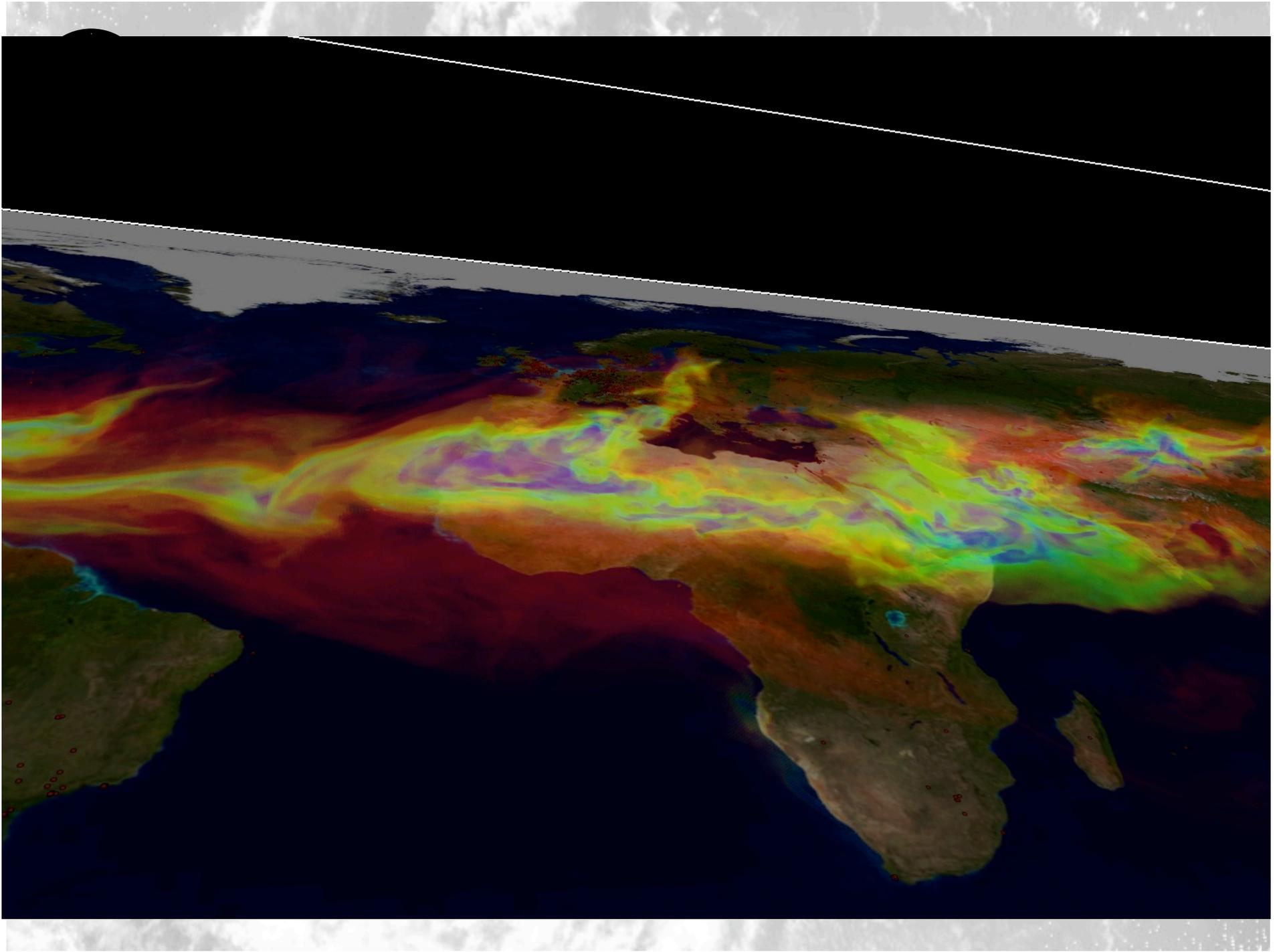
Probe Mode: To modify probe in scene, grab handle with left mouse to translate, right mouse to stretch

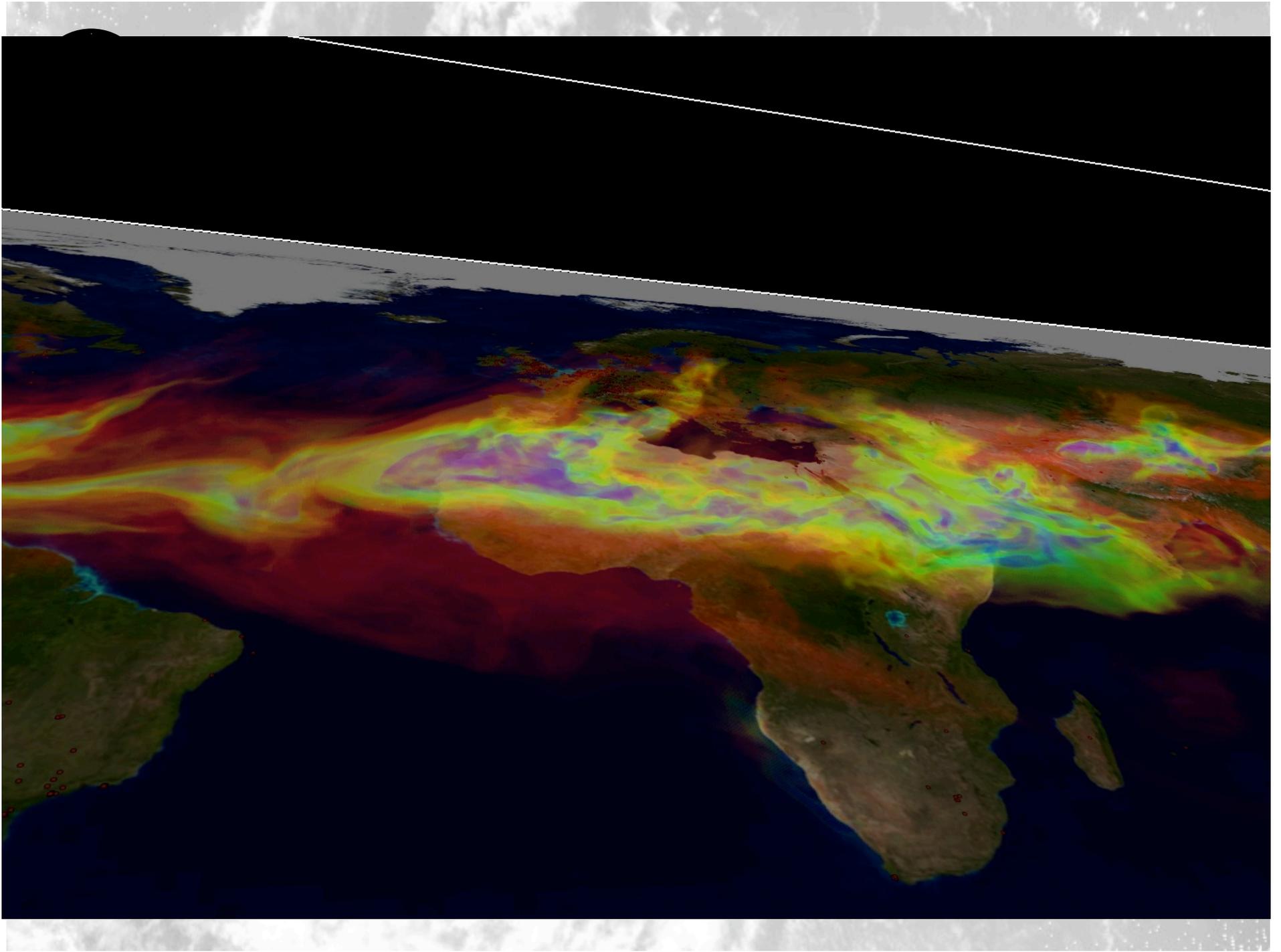


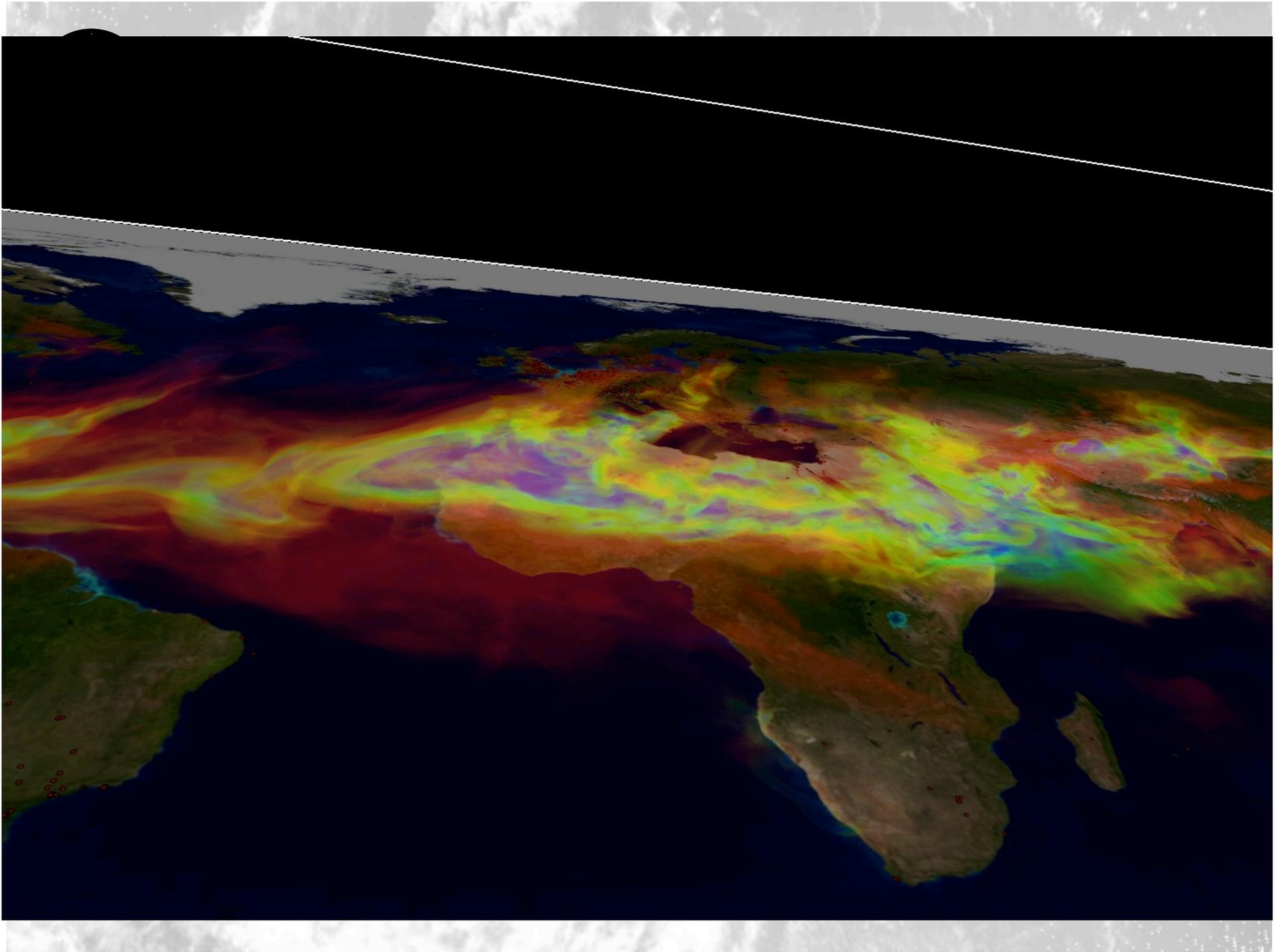


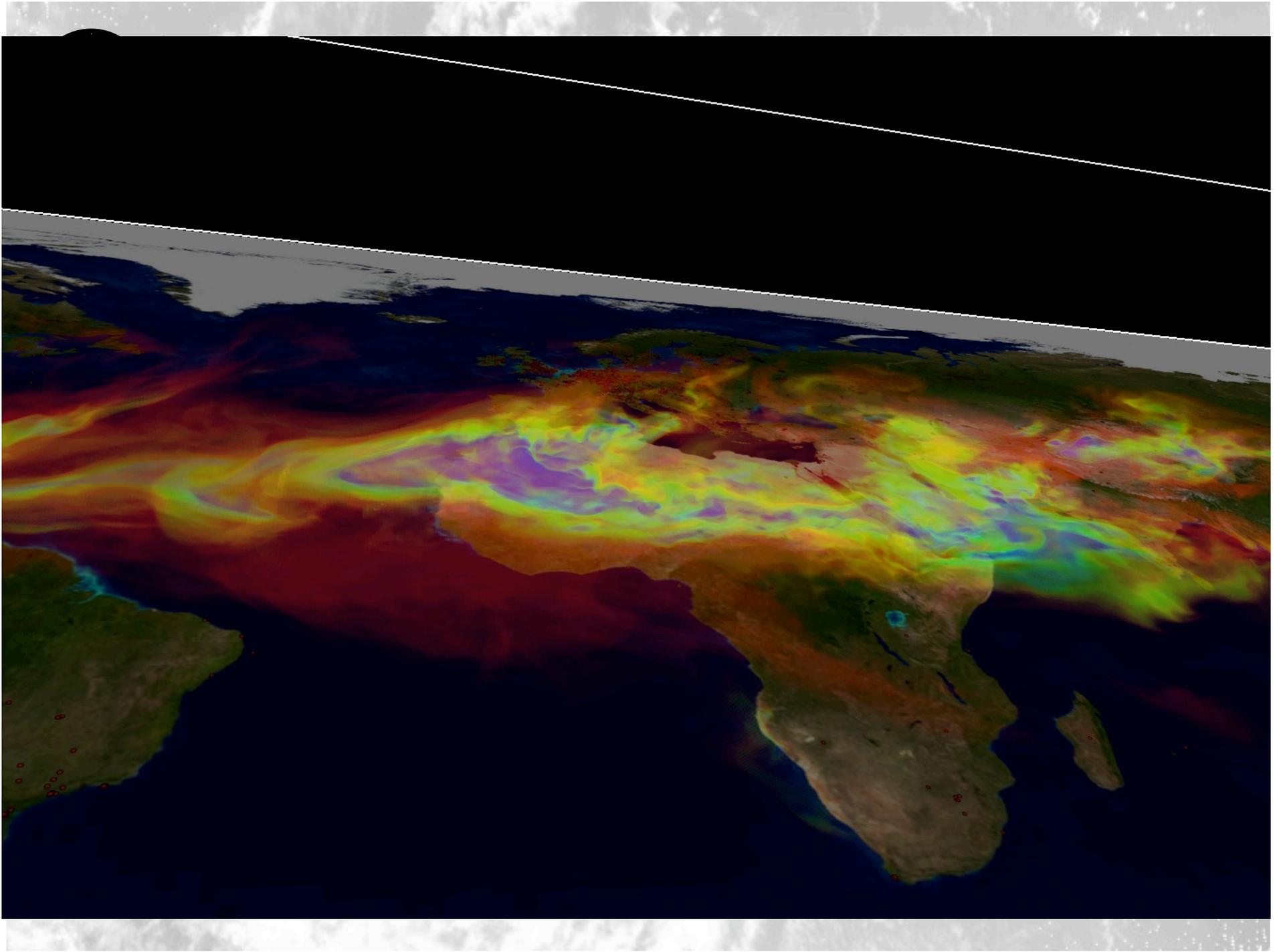


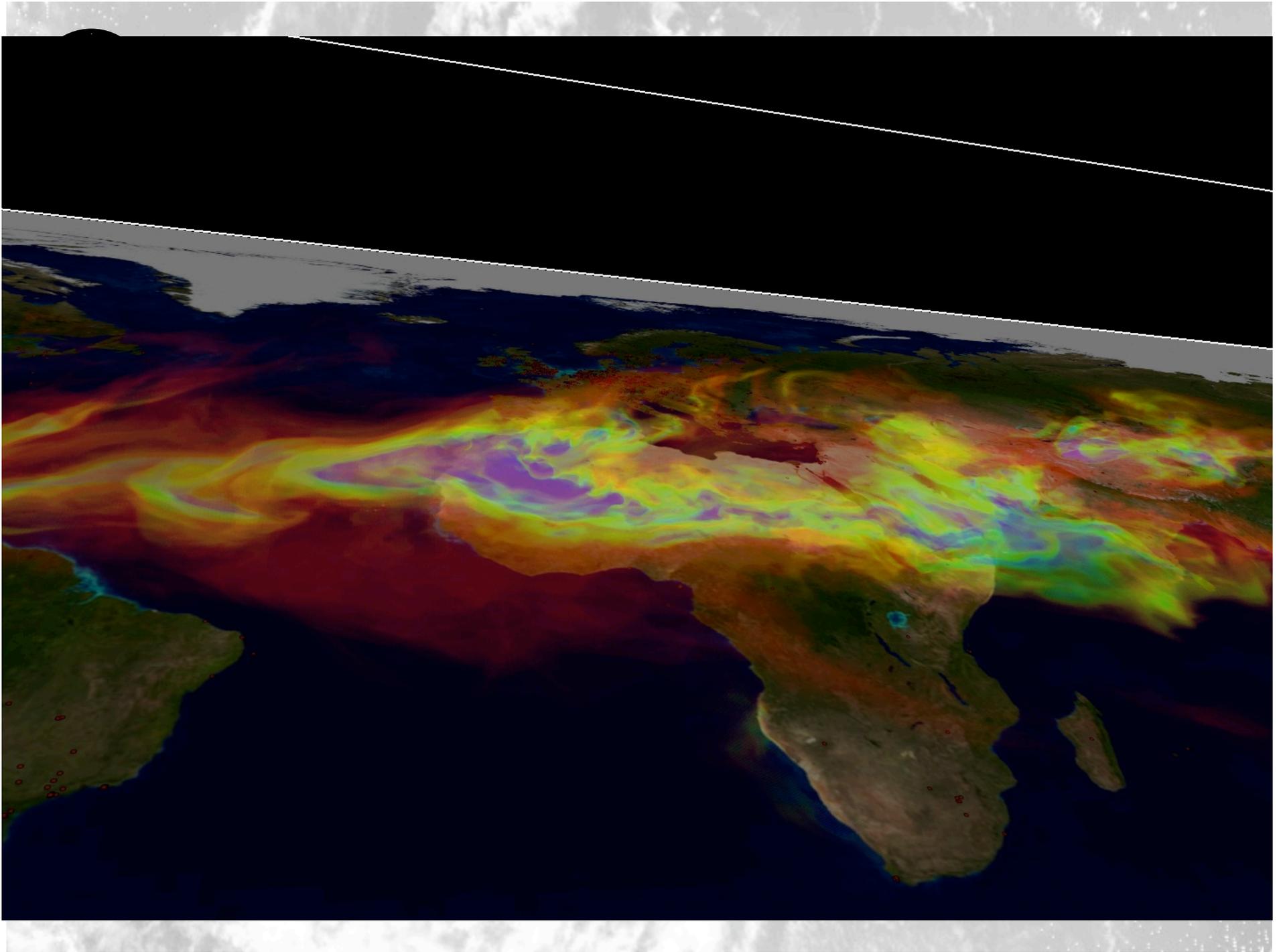


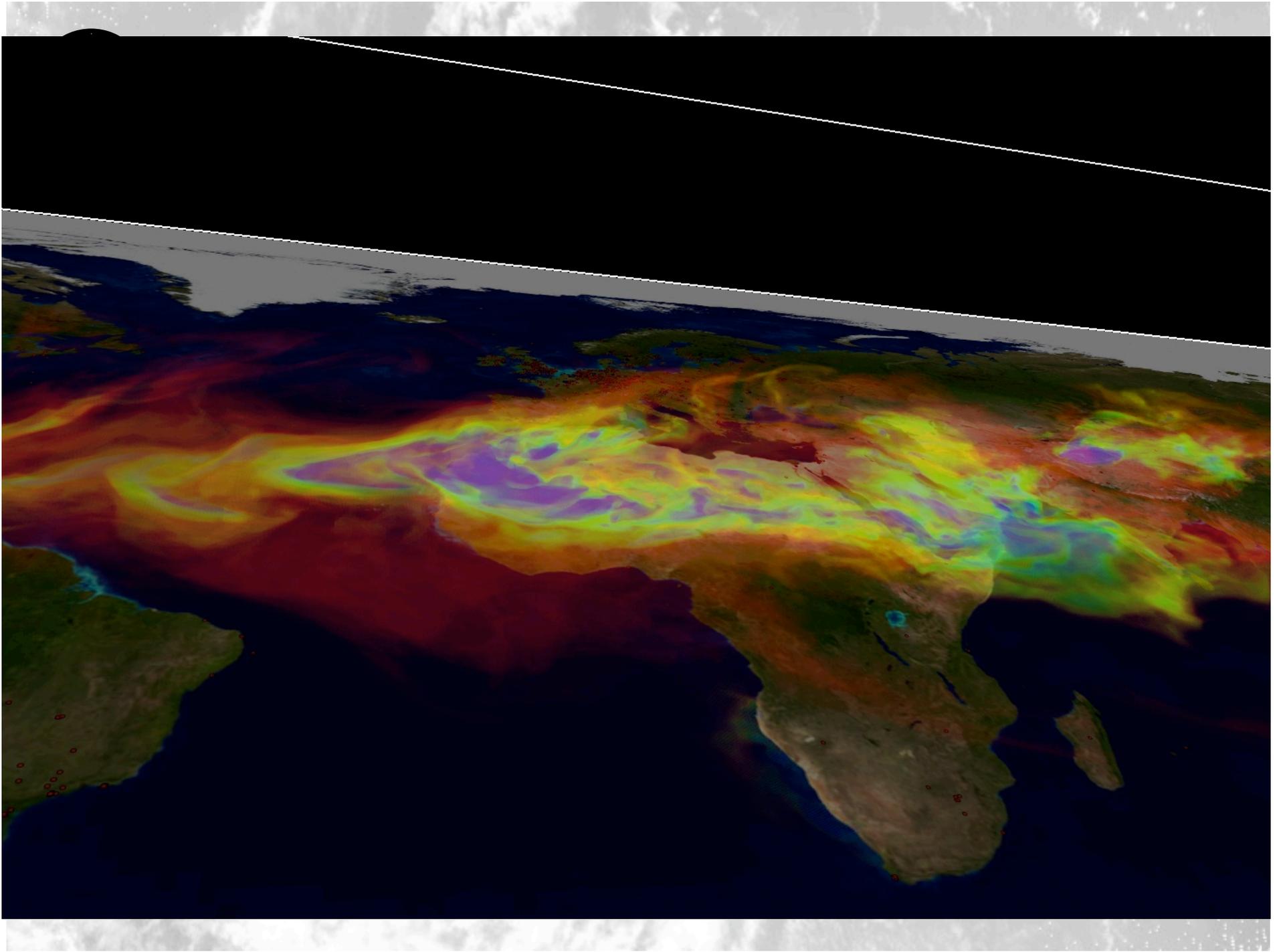


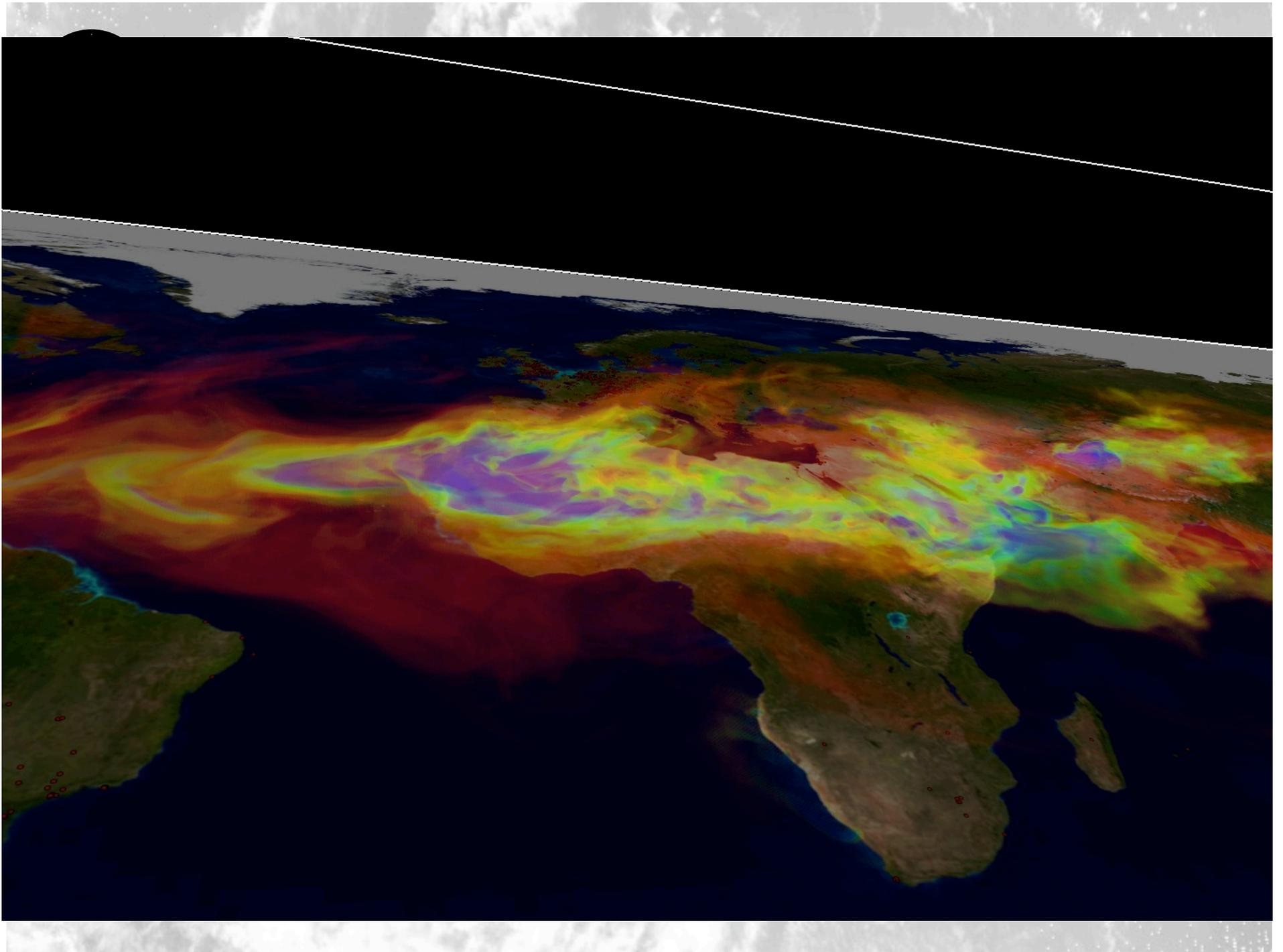


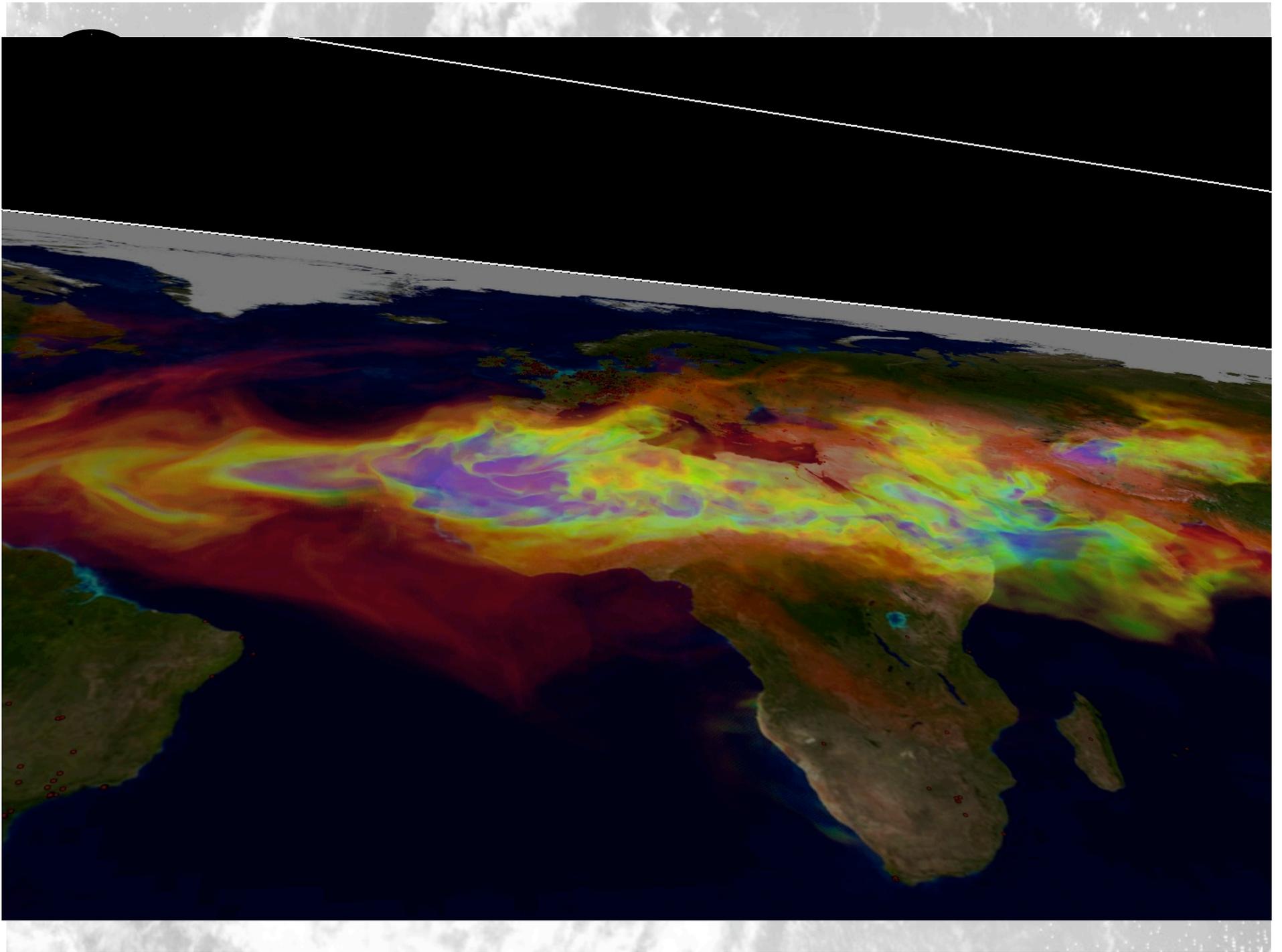


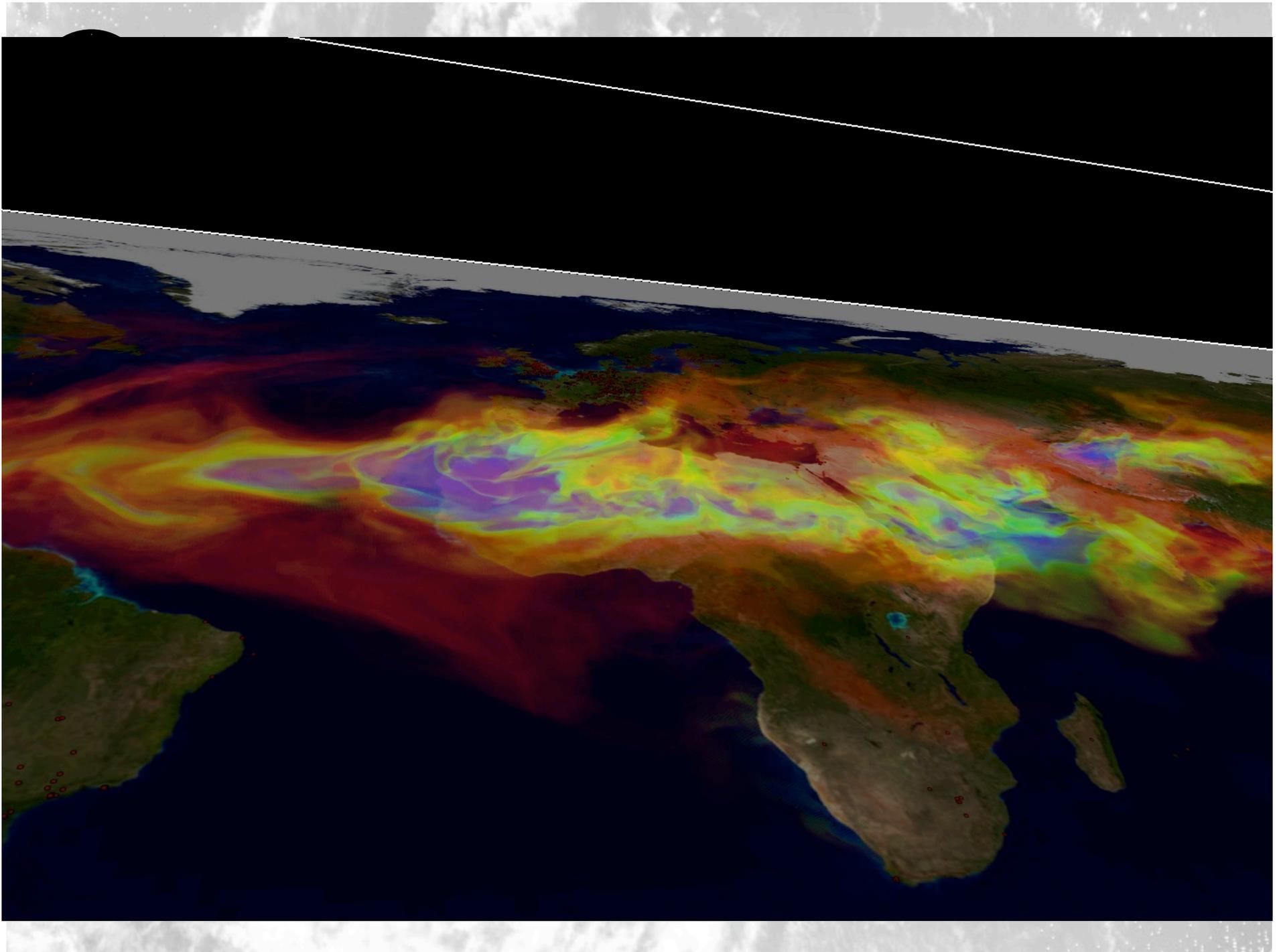


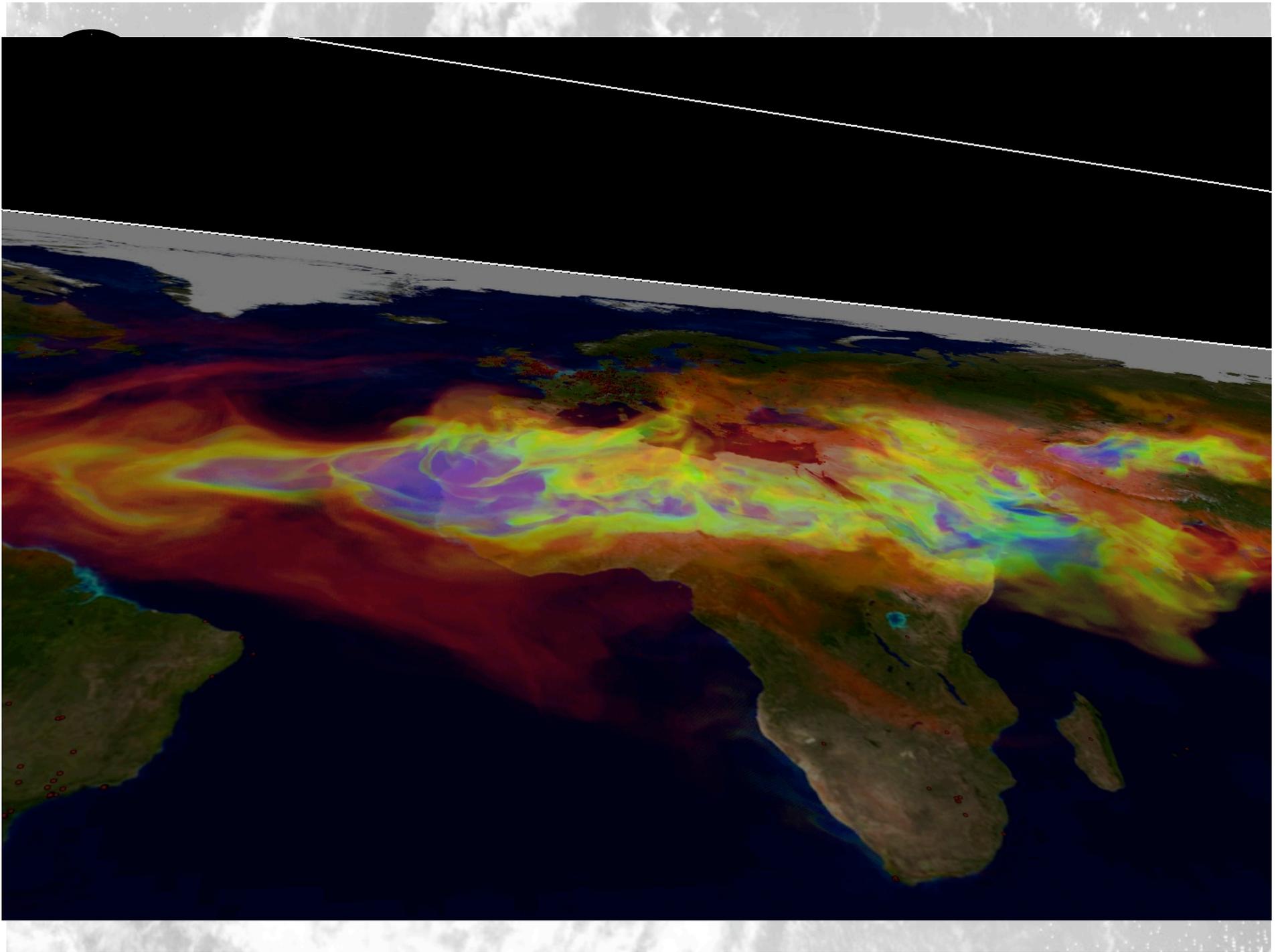


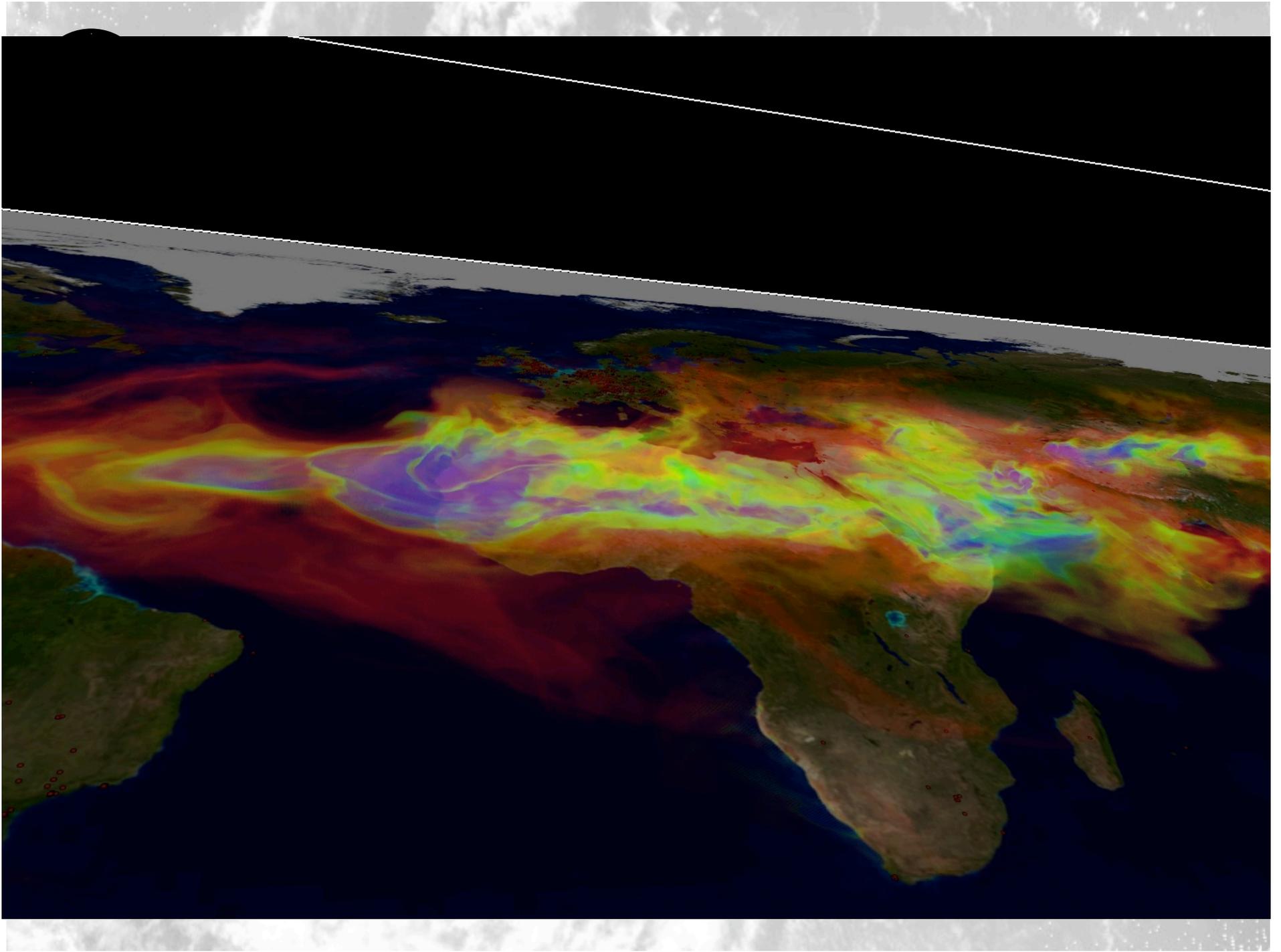


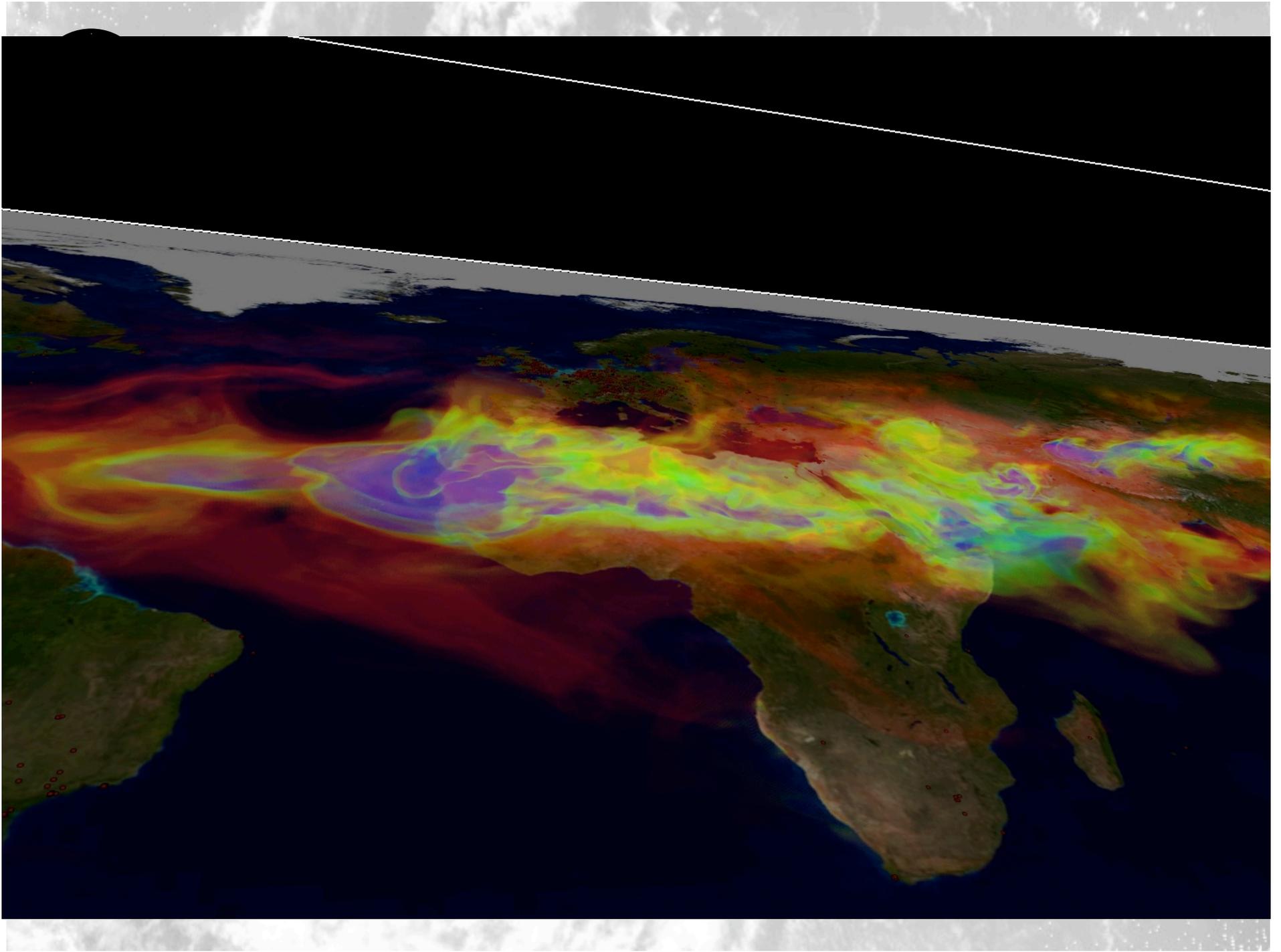


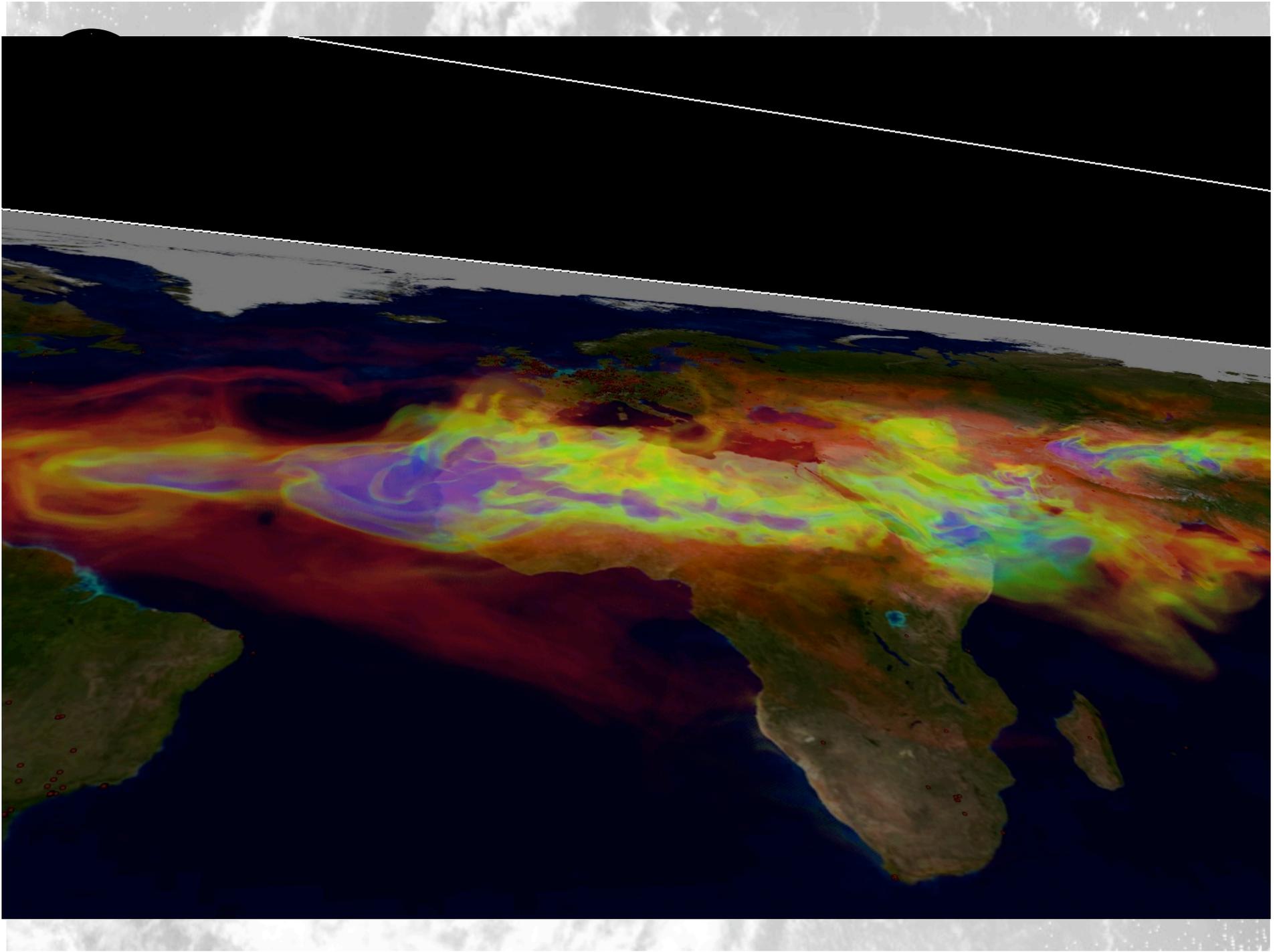


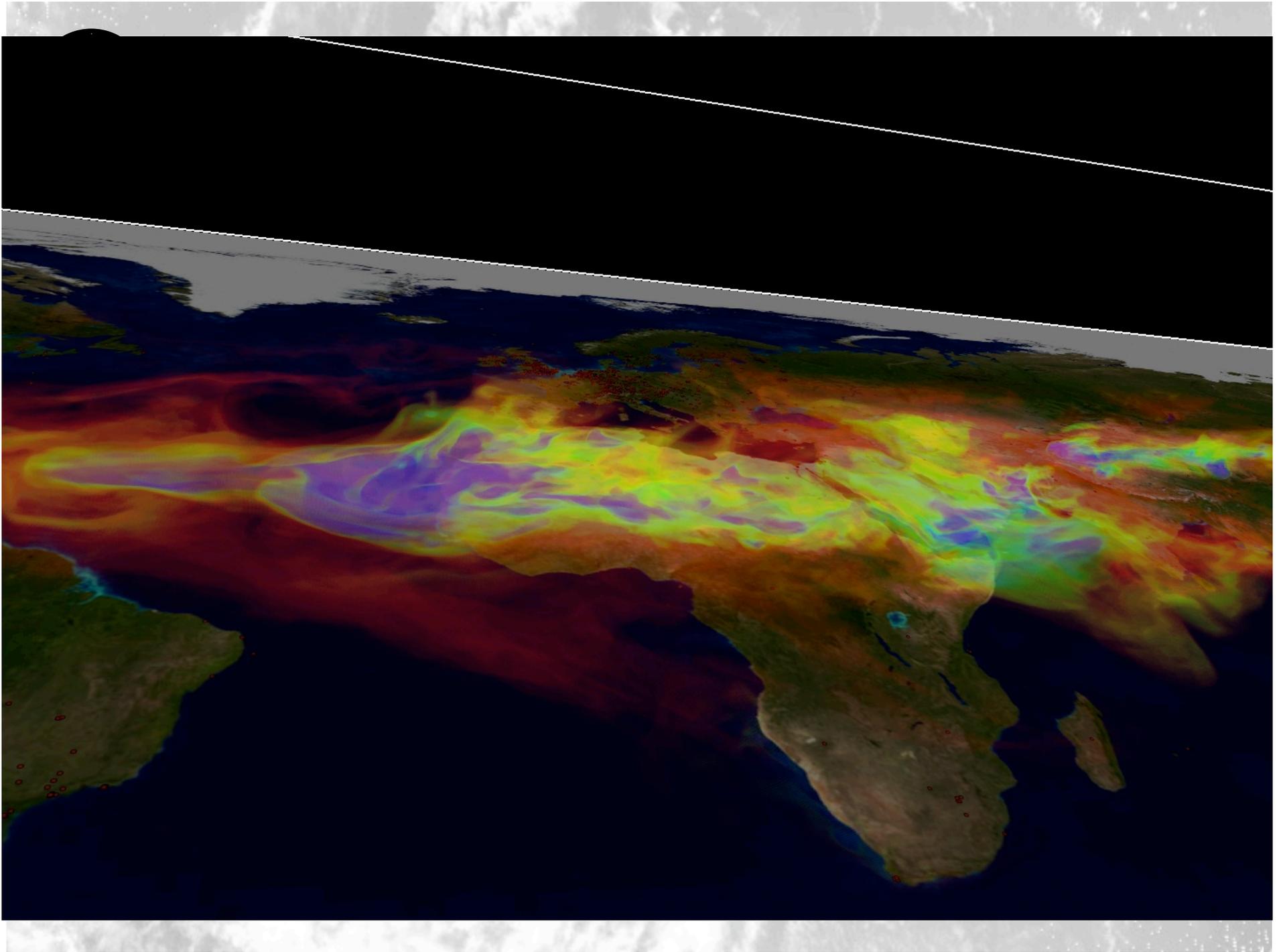


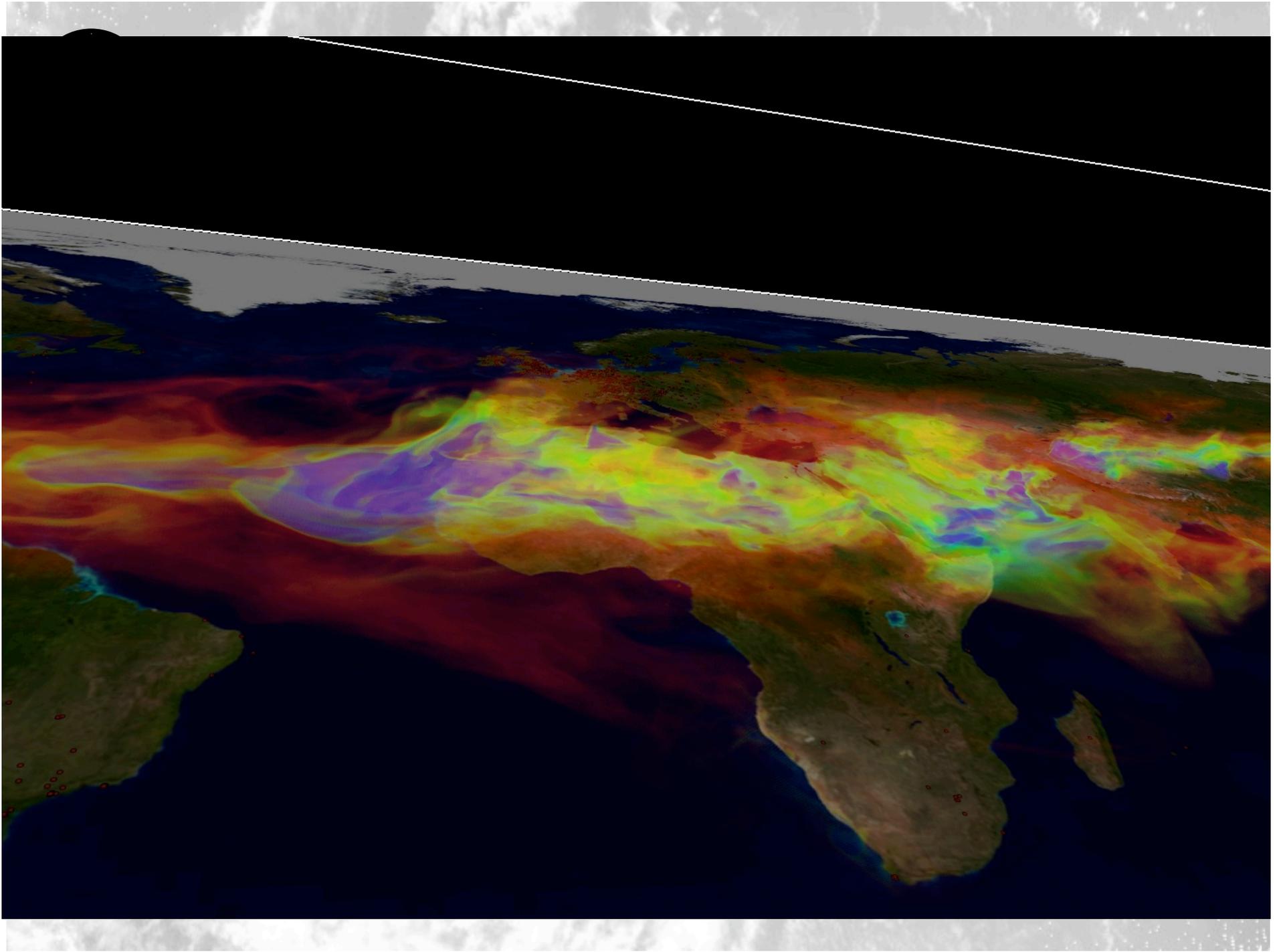


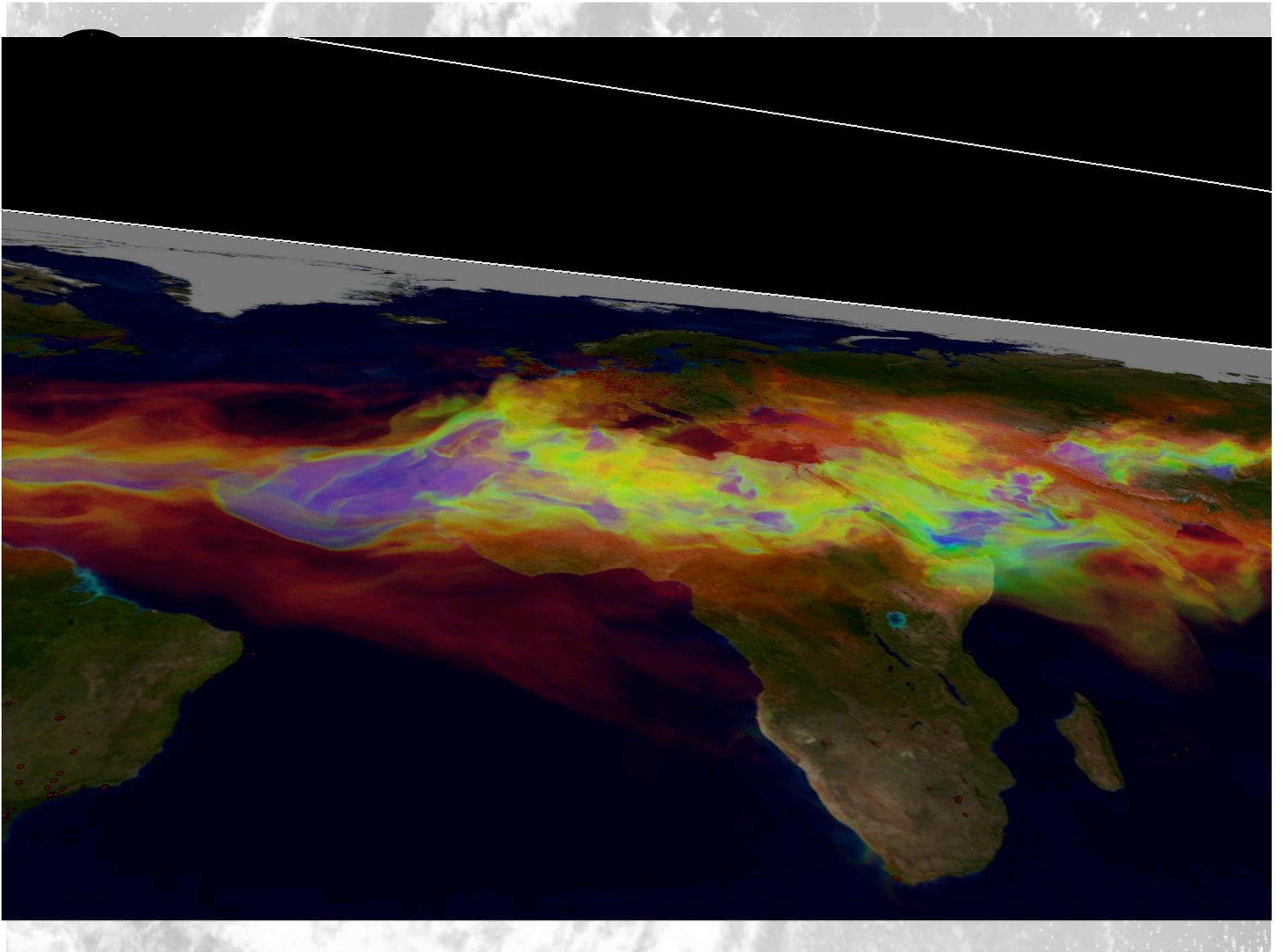














NCCS

# DV3D

- Interactive 3D visualization of simulation data
  - Tailored for climate scientists
- Python app built on MayaVi / VTK
- Simple GUI interface
- Integrated analysis toolkits
  - Runs standalone or as pyGrads (or CDAT) plugin
- Developed at NCCS
  - Adapted to hyper wall and 3D displays



# Agenda

NCCS

Welcome & Introduction  
Lynn Parnell/Phil Webster

User Services Updates  
Tyler Simon, User Services

Current System Status  
Fred Reitz, HPC Operations

Scaling Jobs with MPI on Discover  
Bill Putman, SIVO ASTG

NCCS Compute Capabilities  
Dan Duffy, Lead Architect

Analysis Updates & 3D Demo  
Tom Maxwell, Analysis Lead

Questions and Comments  
Lynn Parnell/Phil Webster



NCCS

# Important Contact Information

NCCS Support:

[support@nccs.nasa.gov](mailto:support@nccs.nasa.gov)

(301) 286-9120