Evaluating a New Analytics Platform Using the Diurnal Cycle of High Frequency Surface Temperature Data

Mughilan Muthupari

Laura Carriere, Gerald Potter, Thomas Maxwell, Daniel Duffy

Science

606.2

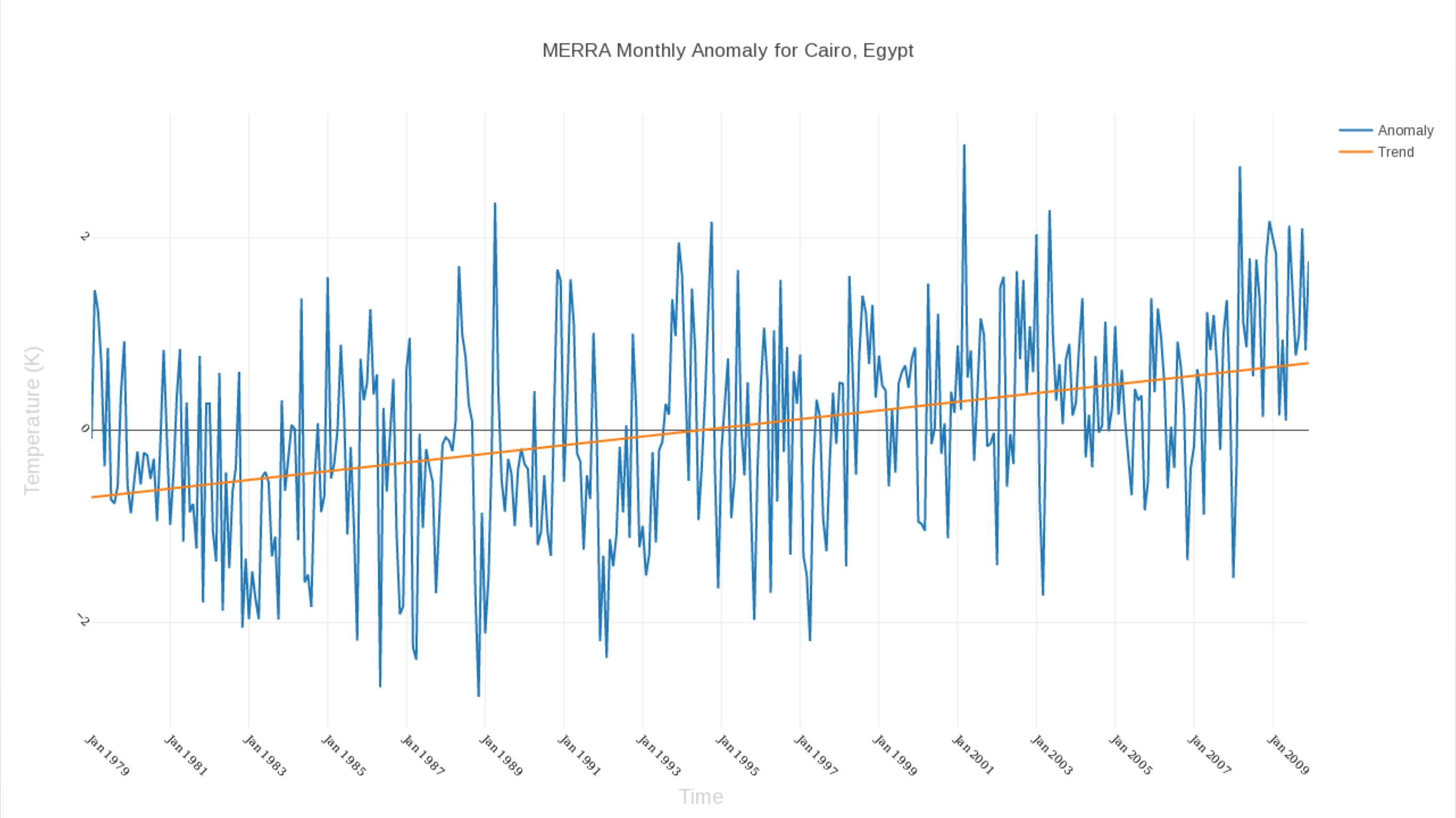
VISGPU02

- One system containing:
  - 2 Intel Xeon E5-2670 Processors
    - 10 Cores
    - 2.50 GHz – 3.50 GHz Turbo
  - 128 GB RAM
  - NVIDIA K5000 GPU
  - HP DL380 G8 Server
- Shared-storage with ADAPT and Discover file systems
- Includes Python and UV-CDAT
- Runs CentOS Linux

Times to Run Tests on Visgpu02 (minutes)

602	705	700	820	715	759

Below: *An example plot produced by the testing, which plots the monthly anomaly for the MERRA dataset for the city of Cairo along with the trend line across the entire time period. The following trend line was significant with a significance level of 0.05.*



ABSTRACT

There are two problems that needed to be solved with this project. One is a big data problem in that we have a huge amount of temperature data that is difficult to handle. NASA has developed a new analytics platform called the Earth Data Analytics System (EDAS) built on a file system, or analytics cluster, called the Data Analytics Storage System (DASS) to handle the data. Theoretically, this is more efficient than NASA's older system, however, we need practical results. This is done by comparing the system currently in use, Visgpu02, and the newer EDAS system to show the difference in the efficiency using various analytics on the data taken from reanalyses. Timings of multiple runs indicate that EDAS is nearly 50 times faster than Visgpu02.

The second goal is to use hourly temperature data to examine the difference in the average diurnal cycle from 5-year time slices from the beginning and end of the available reanalyses to determine if the data is consistent with our understanding of climate change. Multiple cities, one region, and the globe were examined. In the future, these analyses will be placed in Jupyter Notebooks to provide a more interactive experience for the user.

DESCRIPTION OF TESTS RUN

1. Five cities were tested: Moscow, Beijing, Cairo, Phoenix, and Melbourne. The single latitude and longitude coordinate that was closest to each of these cities were calculated and fed in.
2. For Visgpu02, the file was read in using the cdms2 package. For EDAS, the file was read in via a query to the DASS file system. This was done to each reanalysis and all data was saved.
3. Calculate the average of each month's hour (average of all 1 AMs in 1979 Jan, all 2 AMs, etc.) to get an average diurnal cycle for each month in the time period.
4. Calculate both monthly and seasonal composites.
5. Calculate and plot monthly and seasonal anomalies.
6. Perform statistical significance tests on the trend of each anomaly by fitting a line using least-squares linear regression and calculating the significance of the slope value using an F-test for significance.

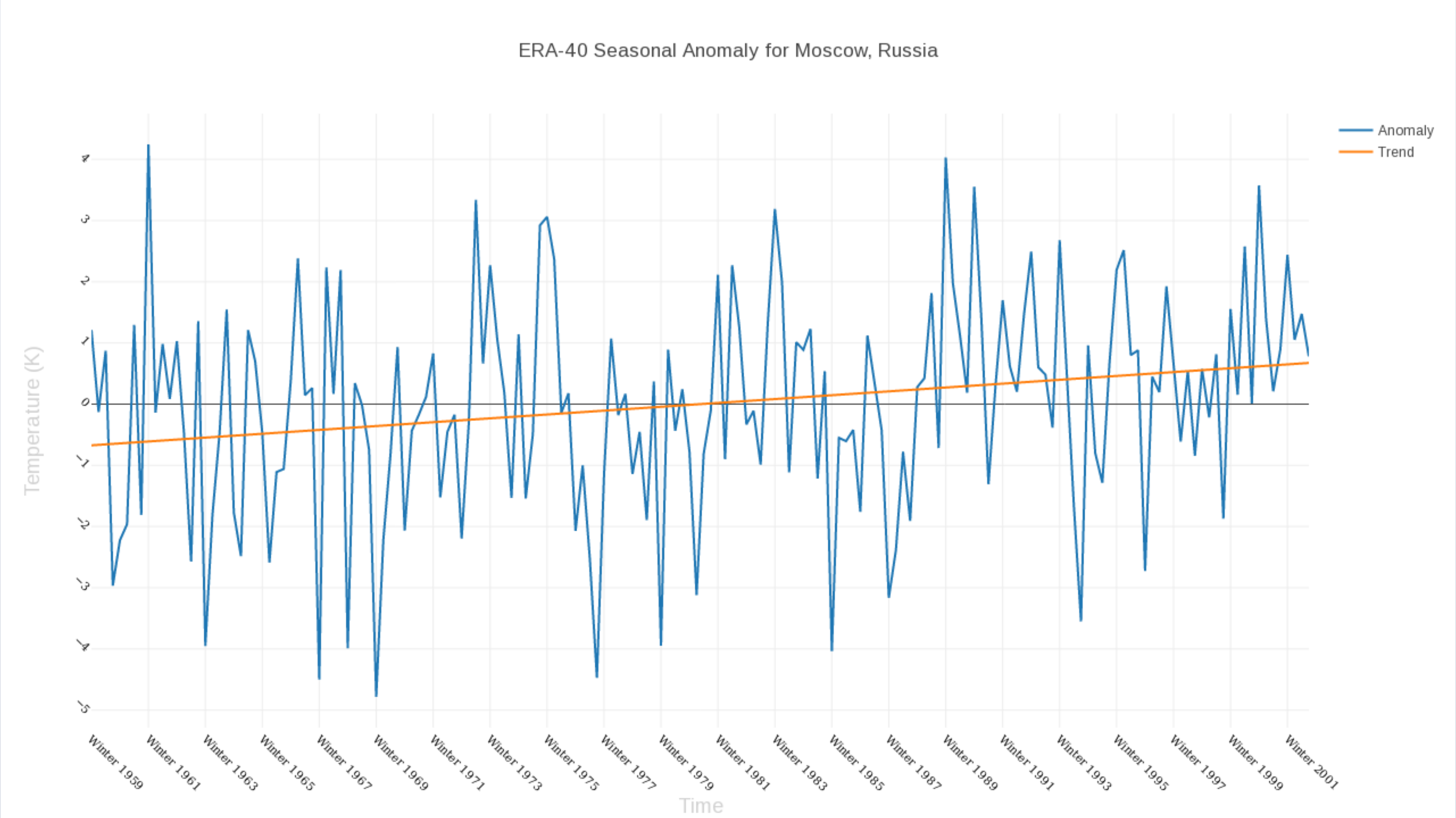
EDAS and DASS

- 15 petabytes raw storage – 12 usable + metadata
  - 1980 8 terabyte disks
  - Metadata on 44 800 gigabyte solid state drives (SSDs)
- 22 nodes configured for persistent virtual machines
- Worked with the POSIX file system formatted with GPFS
  - Allows data to be accessed over multiple connectors at once
- Actual data is stored on spinning drives, but metadata on the SSDs allows for fast discovery
- Runs CentOS Linux

Times to Run Tests on EDAS and DASS (minutes)

14.34	14.20	14.51	14.92	15.22	14.82	14.60	15.21

Below: *An example plot produced by the testing, which plots the monthly anomaly for the ERA-40 dataset for the city of Moscow along with the trend line across the entire time period. The following trend line was significant with a significance level of 0.05.*



INPUT DATA DESCRIPTION

The data comes from four different reanalyses, Modern-Era Retrospective Analysis for Research and Applications (MERRA, 1979-2009), 40-year ECMWF Reanalysis (ERA-40, 1958-2001), Intern ECMWF Reanalysis (ERA-Interim, 1979-2009) and NCEP-NCAR Reanalysis (NRA, 1948-2009), which have been further processed by Wang and Zeng to produce an hourly record. [1]

The data used is hourly surface-air temperature data for the entire time period of the reanalysis at a 0.5° (≈34.5 miles) spatial resolution. This was interpolated using six-hourly data from the above four reanalyses were used: [1]

The interpolation utilized MERRA hourly SAT climatology for each day to go from 6-hourly to hourly along with the Climate Research Unit Time Series 3.10 (CRU TS3.10) to correct the monthly-mean maximum and minimum.

Wang and Zeng found that the final interpolations performed well with respect to real-time hourly data in six cities.

GLOSSARY

Reanalysis: A method for developing a comprehensive record of how climate is changing over time. Both observations and modeled data are combined objectively to generate a synthesized estimate of the state of the system at every point. [2]

Composite: Average across all categories of a specific time period (Ex. average all Januaries, all winters).

Anomaly: Data generated through subtracting each point from the average of all points in that category.

Plotly package: The main plotting package used to plot. It generates plots using HTML, which allows for interactivity in the browser.

Jupyter Notebook: An in-browser development environment geared toward easy presentation of code and output. Allows Python, R, and Scala among others. [3]

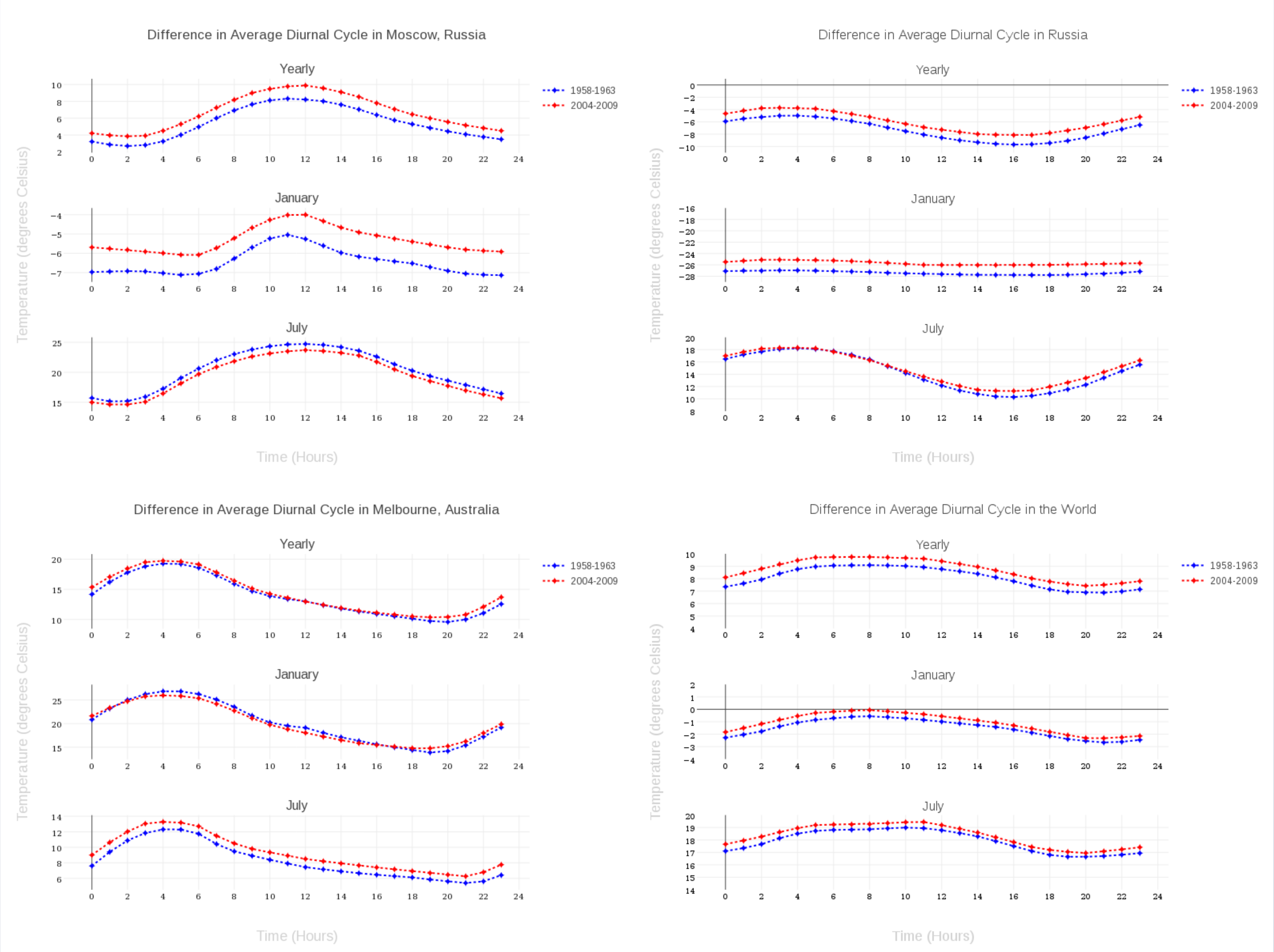
SOURCES

1. Wang, A., & Zeng, X. (2013). Development of Global Hourly 0.5° Land Surface Air Temperature Datasets. Journal of Climate, 26(19), 7676-7691. doi:10.1175/jcli-d-12-00682.1
2. http://reanalyses.org/
3. http://jupyter.org/

ACKNOWLEDGEMENTS

Melissa Cannon, John 'Hoot' Thompson, Laura Carriere, Gerald Potter, Thomas Maxwell, Daniel Duffy

PLOTS



Above: *Plots showing the difference in the average diurnal cycles of the first 5 years (1958-1963) and the last 5 years (2004-2009). NRA and ERA-Interim were available for the first 5, and MERRA and ERA-40 were available for the last 5. The time shown is in Greenwich Mean Time (GMT)*

SUMMARY

Computer Science

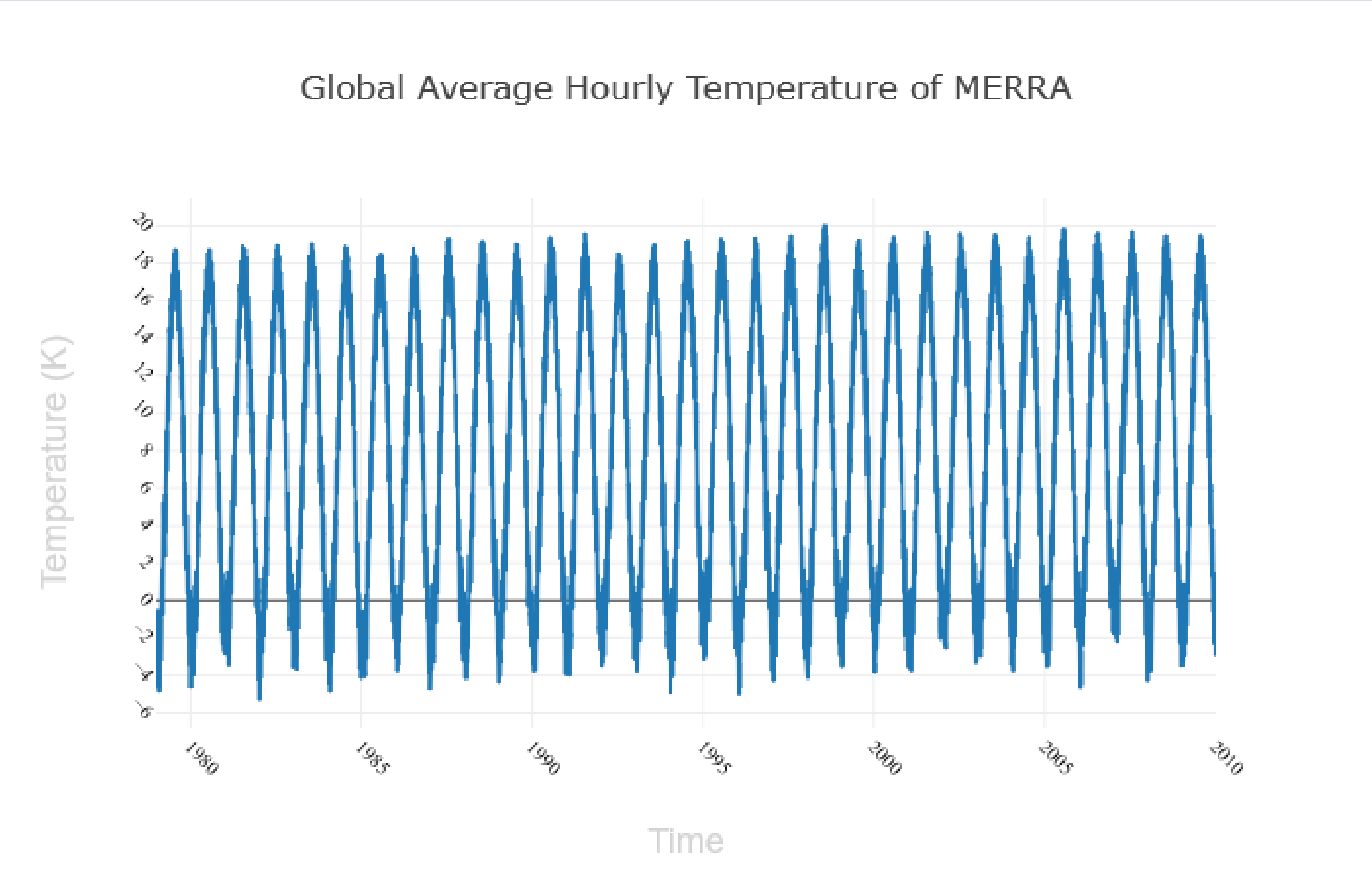
Both Visgpu02 and EDAS/DASS file systems are good for data analytics, however, EDAS is much more efficient in handling very large amounts of data that need to all be processed in a similar way with its metadata stored on SSD and the GPFS architecture. Additionally, it was impossible to perform global calculations on Visgpu02. However, on EDAS, it completed without a hitch. The average time for Visgpu02 was approximately 716.83 minutes, while utilizing EDAS and DASS resulted in an average time of 14.73 minutes.

Science

Using this new data set of historic surface temperature, we wanted to see if the diurnal cycle for the selected cities and regions has changed in some way over the past 50 years. We explored several measures and found all of the cities had significant positive anomaly trend lines for both monthly and seasonal (with the exception of Melbourne) with a significance level of 0.05.Additionally, the graphs on the left show that the cooler months experience a more drastic increase than the warmer months. During the warm months, the minimum temperature has also risen i.e. it doesn't cool down as much during the nighttime hours.

This is consistent to the climate scientists' understanding of global warming.

Below: *The following plot is an example of what is now possible with EDAS and DASS. It averages the temperature of the entire globe at every single hour for the entirety of the MERRA dataset. Visgpu02 would quickly run out of memory if this sort of operation took place.*


Global Average Hourly Temperature of MERRA
Temperature (K)
Time