Title: Cloud Archiving and Data Mining of High-Resolution Rapid Refresh Forecast Model Output

Journal: Computers and Geosciences

Authors: Brian Blaylock[1], John Horel[1], Samuel T. Liston[2]

[1]University of Utah, Department of Atmospheric Sciences

[2]University of Utah, Center for High Performance Computing

Corresponding Author: Brian Blaylock, brian.blaylock@utah.edu

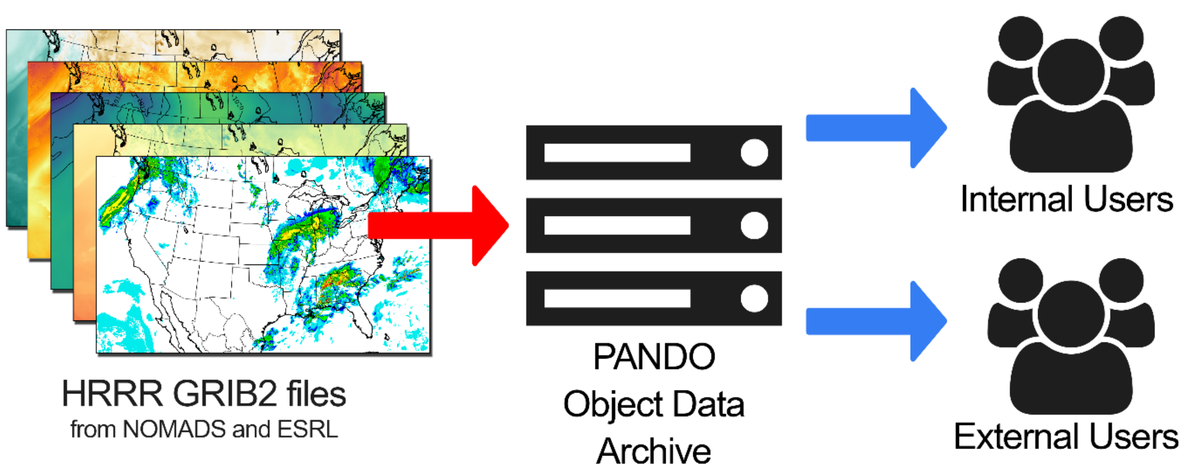
Address: 135 S 1460 E, Rm 819, Salt Lake City, UT 84112

Keywords: object data storage; data stewardship; atmospheric modeling; cloud computing

Abstract

Weather-related research often requires synthesizing vast amounts of data that need archival solutions that are both economical and viable during and past the lifetime of the project. Public cloud computing services (e.g., from Amazon, Microsoft, or Google) or private clouds managed by research institutions are providing object data storage systems potentially appropriate for long-term archives of such large geophysical data sets. We illustrate the use of a private cloud object store developed by the Center for High Performance Computing (CHPC) at the University of Utah. Since early 2015, we have been archiving thousands of two-dimensional gridded fields (each one containing over 1.9 million values over the contiguous United States) from the High-Resolution Rapid Refresh (HRRR) data assimilation and forecast modeling system. The archive is being used for retrospective analyses of meteorological conditions during high-impact weather events, assessing the accuracy of the HRRR forecasts, and providing initial and boundary conditions for research simulations. The archive is accessible interactively and through automated download procedures for researchers at other institutions that can be tailored by the user to extract individual two-dimensional grids from within the highly compressed files. Characteristics of the CHPC object storage system are summarized relative to network file system storage or tape storage solutions. The CHPC storage system is proving to be a scalable, reliable, extensible, affordable, and usable archive solution for our research.

Graphical Abstract



1. Introduction

Weather research and operational weather forecasting depends heavily on evaluating the output from high-resolution regional numerical weather prediction models. The Weather Research and Forecasting (WRF) model is the world's most widely-used regional numerical weather prediction model relied upon operationally for life-saving weather forecasts and for aviation, energy, fire prediction, surface transportation, and water resource management applications (Powers et al. 2017). The High-Resolution Rapid Refresh (HRRR) version of the WRF model, developed by the Earth Systems Research Lab (ESRL), is an hourly updating, cloud-resolving, convection-allowing model run operationally by the National Centers for Environmental Prediction’s Environmental Modeling Center (EMC) (Benjamin et al. 2016). Output from most U.S. operational weather models run by EMC are available on EMC servers for the current day and then archived by the National Centers for Environmental Information (NCEI). However, the voluminous HRRR model output available each hour for forecast durations from 0-18 h with a grid spacing of 3 km over the contiguous United States (1.9 million grid points) is not yet available from NCEI. To archive in a highly compressed format, a representative sample of the output generated by the operational HRRR model requires over 200 TB of disk space per year.

Researchers rely heavily on output from regional models such as HRRR and WRF to diagnose the interplay between complex atmospheric processes on spatial scales from $10^2 – 10^6$ m and temporal scales from $10^2 – 10^7$ s (Benjamin et al. 2016; Powers et al. 2017). A common research strategy is to focus on case studies of specific weather events as a practical approach to manage the TBs of output generated by the models (e.g., Blaylock et al. 2017, Crosman and Horel 2017). With continued growth in computing capabilities, numerical simulations will continue to transition to finer spatial and temporal resolution over increasingly large regional domains. As these models grow, so does the storage space and monetary cost required to archive model output. Of course, large data storage needs are ubiquitous throughout the atmospheric sciences, for example, to archive satellite imagery (Moody et al. 2016) or multi-decadal numerical simulations of the climate system (Taylor et al. 2012).

Molthan et al. (2015) highlight that cloud computing resources (computational services delivered over networks) are providing new capabilities for supporting numerical weather prediction and are a potential solution to archive large volumes of data (Armbrust et al. 2010; Sandholm and Lee 2014). To meet these needs, Sandholm and Lee (2014) described how these services need to be: scalable; fault-tolerant; reliable; high-performance; and easy to use, manage, monitor, and provision efficiently and economically. Public cloud services provided by corporations (e.g. Amazon, Google, or Microsoft) or research consortia (e.g. Open Science Data Cloud, https://www.opensciencedatacloud.org/) are increasingly viable options to meet those requirements, although understanding the extent to which they are economical can be difficult (Chou 2015; Amazon Web Services 2017a). Private cloud services are defined as being operated by an organization for which hardware, networking, storage, and other infrastructure are not directly shared with other organizations (Mell and Grance 2011). The Center for High Performance Computing (CHPC) at the University of Utah provides private cloud services through a data center located off campus.

The objective of this paper is to illustrate the utility and cost effectiveness of a PB disk-based object storage data system managed by the CHPC for archiving large data sets. The capabilities of object data storage systems for geoscience applications will be illustrated in terms of an archive of operational and experimental forecasts from the HRRR model in the contiguous

United States and Alaska from early 2015 to the present. While we have relied extensively over the years on other CHPC storage media (such as a robotic tape archive system and over 100 TB of network file system disk storage), the object data storage system is meeting several of our interwoven needs that are less practical using other traditional data archival approaches: (1) efficient expandable storage for thousands of large data files; (2) data analysis using fast retrieval of user selectable byte-ranges within those data files; and (3) the ability to have the data publicly accessible to the atmospheric science research community.

The remainder of the paper describes how the archive is built and how users can access the data (section 2), followed by applications for which data from the HRRR archive have been used (section 3), and concludes with a discussion of the growing need for large archives and some limitations that should be resolved in the future (section 4).

2. Methods

2.1 *Pando Object Storage System*

The CHPC has dramatically increased its network file system data storage capabilities over the past 10 years from ~400TB to ~14PB due to decreased hardware costs and development of cost-effective storage solutions (Center for High Performance Computing 2017). However, archival storage capacity primarily in terms of a robotic tape system has not increased as rapidly, leaving a large fraction of the data without backup. To help mitigate this shortcoming, CHPC developed a disk-based object storage solution referred to as Pando (named for a vast stand of aspen trees in Utah that is thought to be the largest and oldest single living organism). Currently at 1 PB in usable capacity, Pando was developed at lower cost than other archival options and has greater resiliency, accessibility, and expandability. Researchers lease dedicated amounts of archival space over a 5-year span to help recover some of the costs for Pando. They then manage their own space, which helps reduce CHPC's administrative burden to manage the archive.

The CHPC took into consideration that an improved archival system needed to scale to a much larger size than what might be affordable initially. Large network file systems or Redundant Array of Independent Disks (RAID) sets do not scale well as the number and size of drives increase, particularly since recovering and repairing after an error or disk corruption may require disks to be offline for many days. The CHPC selected Red Hat's Ceph object-based open source storage system (Maltzahn et al. 2010) to address the shortcomings of both RAID and file systems based on published performance comparisons (e.g., Poat et al. 2015) and testing over several years. Low-level operations, such as block or file level I/O, are managed by a software layer that manipulates objects for the user or administrator such that expensive RAID controllers are not necessary and archived objects can be replicated or made redundant according to configurable parameters.

Pando was formatted using the 6+3 erasure coding, i.e., all objects are broken into 9 pieces—6 data pieces and 3 redundancy pieces necessary for data protection and reconstruction. The initial 1 PB Pando archive consists of 9 storage servers each with sixteen 8 TB drives that are coordinated by 3 monitor nodes that efficiently maintain the map of the objects in the system (Fig. 1). If the file system on a single drive becomes corrupt, then: (1) that drive is logically removed by the system administrator; (2) the administrator recreates the file system and logically adds it back in; and (3) the objects are redistributed within the new file system automatically by the Ceph software to maintain the configured level of redundancy. The 6+3 erasure coding ensures no data loss even if every disk fails on three servers. The Pando system has the capacity to contain 44 servers before additional network infrastructure must be purchased making it expandable to approximately 5PB with current drive capacities. To ensure that Pando is in production past disk warranty periods, Ceph can transparently migrate the data to new hardware when old hardware is retired.

The Amazon Simple Storage Service (S3) has been implemented on Pando through a Reliable Autonomic Distributed Object Store (RADOS) Gateway node to focus on long-term storage needs separate from the other mounted file systems available to CHPC users (Nawaz et al. 2016). The RADOS Gateway node (Fig. 1) serves as an interface between client computers and objects managed by the RADOS software layer. Present usage suggests that additional RADOS Gateway nodes will be necessary in the future to avoid throughput bottlenecks (speeds of only 5 GB s^{-1} during high loads) that limit optimal utilization of the Pando system. Objects are most efficiently uploaded to Pando from the CHPC local file systems using rclone (Wood 2017), which is open source software commonly used to download or upload files between hard disk and cloud storage systems.

2.2 *HRRR Data Archive*

Several implementations of the HRRR modeling system have been developed by ESRL researchers with staff at EMC maintaining its operational version for the contiguous United States (Benjamin et al. 2016). To support air quality research at the University of Utah (Horel et al. 2016; Blaylock et al. 2017), we started archiving operational HRRR analysis (forecast hour 0) output files beginning April 2015 on local network file system disks obtained from the NOAA Operational Model Archive and Distribution System (NOMADS). Other research projects led us to download selected meteorological fields from the operational HRRR 1-18 h forecast files beginning in summer 2017 and analysis and forecast fields from experimental versions of the HRRR for the contiguous United States and Alaska. The thousands of 2-dimensional meteorological fields available from the HRRR are stored as gridded binary-2 (GRIB2) files, a highly efficient binary format that relies on Joint Photographic Experts Group (JPEG) 2000 image compression (Silver and Zender 2017).

By early 2017, local file system storage for the HRRR products grew to over 20 TB with the expectation that by later in 2017, over 100 GB of model grids would be added per day. That storage approach was becoming unwieldy to manage across multiple file server partitions and not practical to facilitate access to the archive for an increasing number of atmospheric science researchers external to the University of Utah, who became aware of it through online searches for HRRR model output. After initial testing of the Pando system, all the locally-archived HRRR files were transferred to it and removed from the local file system.

Since EMC and ESRL provide efficient access for anyone interested in HRRR model output for the current and previous day (Bowman and Lees 2015), we prefer external users to not overwhelm our archival system by requesting what is already easily available from those sources. We execute download scripts after 00 UTC to retrieve files for the previous day to our local CHPC network file storage, a process that can take several hours to complete even with multithreading. The files are then copied to the Pando archive using the open source rclone utility. The s3cmd utility is used to change permissions for each file from private to public so they can be accessed by other researchers at the University of Utah and elsewhere.

The present implementation of Ceph on Pando limits the ability to view the contents or manipulate the data object files. Rather, each file has a unique URL that can be used to download it via HTTPS. While anyone can attempt to directly download such files from the archive, web pages have been developed for identifying which files are available to simplify interactive downloads (https://hrrr.chpc.utah.edu; Fig. 2). Users are encouraged to avoid excessive reliance on the interactive pages and create automated download procedures using wget or cURL with example code provided on the aforementioned web page.

Since most users prefer to access a relatively small number of the meteorological fields contained within each of the large HRRR GRIB2 files, it is cumbersome to retrieve the entire file and then process it to extract the fields of interest. To facilitate access to specific 2-dimensional fields, we use the wgrib2 tool (Climate Prediction Center 2017) to create a metadata file for each GRIB2 file and provide that information on a local web server since there is no need to store them as objects in Pando. These index files contain for each field its abbreviated variable name, vertical level, beginning byte, time of the model run, and forecast hour. Hence, it is straightforward to derive the corresponding byte range for a variable and retrieve using cURL its 2-dimensional field. Unfortunately, it is not currently possible to retrieve a byte range within a GRIB2 formatted file for a subsection of the two-dimensional grid (e.g., for a state or regional area of interest). This is a present limitation of object storage and GRIB2 file formats that may be solved through continued development of object storage systems or archiving the gridded data in a different file format. Hence, the smallest granule that can be retrieved from a GRIB2 HRRR file is a single field over that entire domain (~1 MB). Multiprocessing and multithreading techniques such as those available using Python's multiprocessing module can be leveraged to spread the work across multiple cores and reduce download time and greatly increase the data processing speed when fields from multiple files are needed. We have developed Python multi-processor procedures that rely on basic cURL commands to efficiently access the HRRR files from a single dedicated CHPC server. For example, computing the minimum, mean, and maximum wind speed from nearly 17,000 hourly analyses at the 1.9 million grid points in the operational HRRR model was done in less than 15 minutes using 30 processors.

The current HRRR archive directory tree for both the Pando and metadata archive is branched by model type (operational HRRR, experimental HRRR, and experimental HRRR Alaska), by file type (sfc files contain a selection of 2-dimensional fields while many more 2-dimensional fields at fixed pressure levels in the vertical as well as other levels are available in the prs files), and by date (year, month, and day).

```
HRRR/
   ↳ oper/
        ↳ sfc/
             ↳ YYYYMMDD/
        ↳ prs/
             ↳ YYYYMMDD/
   ↳ alaska/
        ↳ sfc/
             ↳ YYYYMMDD/
        ↳ prs/
             ↳ YYYYMMDD/
   ↳ exp/
        ↳ sfc/
             ↳ YYYYMMDD/
```

Each file within the daily directories follow the same naming convention used by NOMADS when the file is first downloaded (files from ESRL are renamed to match the NOMADS naming convention). The files are named by the model type, the initialization hour, variable field, and the forecast hour ([hrrr/hrrrAK/hrrrX].t[hour]z.wrf[sfc/prs]f[forecast].grib2). For example, the following request will download the full surface field file from the operational HRRR analysis for 14:00 UTC 5 April 2017:

https://pando-rgw01.chpc.utah.edu/HRRR/oper/sfc/20170405/hrrr.t14z.wrfsfcf00.grib2.

Metadata for the corresponding HRRR file can be found in the GRIB2 index file located here:

https://api.mesowest.utah.edu/archive/HRRR/oper/sfc/20170405/hrrr.t14z.wrfsfcf00.grib2.idx.

The index file can be used to request specific variables within a byte range. If a user was only interested in 10 m gusts, then the index file indicates that the byte range for the gusts variable for that file is between 3478099 and 4879421. Using cURL, a user can download the gust variable from the larger file as follows:

```
curl -o downloaded_file.grib2 --range 2757386-4110515 https://pando-
rgw01.chpc.utah.edu/HRRR/oper/sfc/20170405/hrrr.t14z.wrfsfcf00.grib2.
```

3. Applications

3.1 High-Impact Weather Events

While voluminous sets of graphics of analysis and forecasts fields from the HRRR model runs are generated routinely by ESRL, EMC, academic institutions, and commercial sources of weather information, those usually depict only conditions within the past few days and only show a small fraction of the information contained in the HRRR GRIB2 files. The HRRR Pando archive provides users access to all the fields contained in the HRRR grib2 files. These files can be used to create customized graphics of high impact weather events or other features of interest to the user. For example, the major New England snowstorm on 14 March 2017 is depicted by the HRRR mean sea level pressure analysis valid at 1700 UTC 14 March 2017 (Fig. 3).

Hourly changes in atmospheric conditions at specific locales can be examined by downloading the requisite grids each hour, which can be easily retrieved from the Pando archive using the procedures described above. Figure 4 illustrates the conditions analyzed by the HRRR centered on 2100 UTC 27 April 2017 at which time a wildfire near O'Donnell Texas traversed across the site of a West Texas Mesonet station (Schroeder et al. 2005) as evident by the 58^{o}C observed 2-m air temperature at that time. The HRRR hourly analyses closely track observations (albeit not the temperature spike associated with the fire) as well as provide additional diagnostic variables, such as winds at 80 m above ground level and estimates of the boundary layer depth.

Since the primary purpose of the operational HRRR model is to provide short-term (0-18 h) weather forecast guidance updated every hour to predict severe weather (Benjamin et al. 2016), assessing the model's ability to properly forecast such conditions is of high interest. For example, 30 tornadoes and hundreds of reports of hail and high winds were received on 4-5 April 2017 from Missouri to Ohio extending southward to Alabama and Georgia (Storm Prediction Center 2017). Airline operations in Atlanta were severely affected on 5 April causing thousands of delayed or canceled flights. Figure 5 contrasts the simulated composite reflectivity and gust analyses from the HRRR model at 1400 UTC 5 April 2017 to the 16 h forecast from the HRRR run initialized 2200 UTC 4 April 2017. The model forecast at 16 h highlights many of the locations that later received heavy precipitation and strong winds.

3.2 HRRR Model Composites

Statistics derived over long-time intervals from model output can provide useful information, such as availability of wind and solar energy resources (James et al. 2017) or identifying model performance characteristics (Katona et al. 2016, Ikeda et al. 2017). Preliminary basic statistics (minimum, mean, maximum, and percentiles) of meteorological variables (temperature, wind speed, snow cover, lightning, etc.) have been derived from the 2-year archive of HRRR analysis grids. Multiprocessing techniques were used to speed up downloading the files from the archive and processing the grids for each of the 1.9 million grid points. Figure 6 shows the 95th percentile of the 10 m gusts analyzed by the operational HRRR at 2300 UTC during all days between 18 April 2015 and 30 March 2017. Such statistics are intended to be used to provide realistic bounds for observations of wind and other variables at over 25,000 locations in the United States that are available within the past 20 years as well as received continuously as part of the MesoWest and SynopticLabs projects (Horel et al. 2002; SynopticLabs 2017). Simultaneous calculations that require less memory (e.g., extreme and mean values) were completed in about 15 minutes for one variable over the entire contiguous United States. Brute-force approaches to calculate multiple percentile values (e.g., 1st, 5th, 10th, 90th, 95th, and 99th) for each hour of the day necessary to generate Figure 6 required storing more values in memory and required roughly an hour for a single variable. Improved approaches using approximation techniques are possible to efficiently compute percentiles and other statistics and avoid excessive memory consumption on our compute nodes.

3.3 Initializing WRF Simulations

The original impetus for our archive of the HRRR output was to obtain the best possible high-resolution WRF simulations over northern Utah to understand a poor air quality episode in the vicinity of Salt Lake City during 17-18 June 2015. Blaylock et al. (2017) ran a 1 km WRF simulation for northern Utah with initial and boundary conditions obtained from the HRRR hourly analyses beginning at 0000 UTC 14 June 2015 and continuing until 0700 UTC 19 June 2015.

While many researchers initialize high-resolution model simulations from operational and reanalysis modeling systems (e.g., Foster et al. 2017; Li et al. 2017), the HRRR provides significant advantages in terms of its 3 km grid spacing, hourly output files, and advanced data assimilation techniques. To the best of our knowledge, the study by Blaylock et al. (2017) was the first one to use HRRR analyses to initialize and provide the requisite lateral boundary conditions for WRF research simulations. While ESRL maintains an internal tape archive of HRRR model output, the HRRR archive on Pando is currently the only readily available resource for other researchers to initialize high-resolution WRF simulations with HRRR boundary conditions. While it is recommended to initialize WRF simulations with native or model-level HRRR files, we don't archive the native level files at this time due to its large file sizes (> 600 GB per file). However, WRF can be initialized with the HRRR pressure-level analysis files available on Pando. The steps required to initialize WRF with HRRR boundary conditions have been documented by Blaylock (2017).

4. Discussion and Conclusions

The management and distribution of large geoscience data sets have received increasing attention, particularly given the explosion in public and private cloud-based resources. For example, an Amazon Web Service (AWS) S3 object store hosts the level 2 retrospective and real-time archive of Next Generation Weather Radar (NEXRAD) data (Amazon Web Services 2017b). Our research group in the Department of Atmospheric Sciences uses Amazon AWS including its S3 object store for other applications that require uninterruptible computational resources and require a relatively fixed small amount of disk storage (SynopticLabs 2017). The complexity and volatility in the egress costs to upload or download data depending on the policies of each public cloud storage facility precluded our use of one of them for the HRRR archive.

The private cloud CHPC Pando object storage archive has made it possible to efficiently archive, access, and analyze the HRRR model output. Pando is also being used by other atmospheric scientists, anthropologists, geneticists, and cancer researchers at the University of

Utah. Our HRRR archive has many of the properties of an ideal data archive described by Kruger et al. (2006)—it is scalable, extensible, inexpensive, and usable. Having fixed leasing costs over a 5-year period allows us to plan as our archival needs grow. The private cloud Pando system provides faster access to our long-term data archive for our needs as well as provide reasonable access times for the several dozen researchers outside the University of Utah that have already discovered its utility in the short time that the archive has been available.

The major limitation of the present Pando object storage systems is that Ceph constrains how the objects can be managed and accessed. Red Hat now supports Ceph File System (Ceph FS, Red Hat 2017) as a Portable Operating System Interface (POSIX) compliant file system that is more flexible to handle the objects in the storage cluster. However, S3-type objects still must be downloaded to a local disk before the data contained within them can be processed. To avoid excessive downloading of data not of interest to a user, the highly efficient GRIB2 format of the HRRR model output allows selecting by byte range and returning only the fields of interest from the many two-dimensional fields contained within an object. Other file formats, such as Hierarchical Data Format Version 5 or Network Common Data Format, may eventually allow subsetting of S3 objects by variable, region, single grid point, all vertical levels at a point, etc., but that capability is not presently available.

We expect that NCEI or other government or institutional repositories will begin to archive operational HRRR model output at some point. Although long-term archives of evolving experimental versions of models are seldom undertaken, having the ability as we do to compare output from experimental and operational versions of the same model makes it possible to assess model improvements more efficiently. Research agencies such as the National Science Foundation now require data management plans that describe what will happen to the data and metadata that led to the research results. While a small number of geoscience data repositories exist (e.g., the National Center for Atmospheric Research), those entities have strict standards for accepting large data sets that are often difficult to meet. At the present time, geoscience data journals require that data sets be in such data repositories prior to publication such as that by Jacques et al. (2016). Academic institutions will increasingly need to consider having facilities like the Pando archive to effectively meet those data stewardship requirements. However, it remains unclear whether those institutions are willing to subsidize the cost of maintaining large archives that are necessary to store results once research projects have been completed and funds are no longer available from the granting agencies.

Acknowledgements

We appreciate the support and resources made available by the Center for High Performance Computing at the University of Utah. We would like to thank the model developers at ESRL and EMC for the ongoing development of the HRRR as well as providing the model output. We also would like to thank Chris Galli for his suggestions about this research and comments on the manuscript. Funding: This research has been supported by the National Science Foundation (NSF Grant 1443046) and the NOAA Collaborative Science, Technology, and Applied Research (CSTAR) Program (NOAA Grant NA13NWS4680003).

References

Amazon Web Services, 2017a. Amazon EC2 pricing. URL: https://aws.amazon.com/ec2/pricing/.

Amazon Web Services, 2017b. NEXRAD data archive. URL: https://aws.amazon.com/noaa-big-data/nexrad/.

Armbrust, M., and coauthors, 2010. A view of cloud computing. *Communications of the ACM*, 53, 50-58. doi: 10.1145/1721654.1721672.

Benjamin, S., and coauthors, 2016. A North American Hourly Assimilation and Model Forecast Cycle: The Rapid Refresh. *Monthly Weather Review*, 144, 1669-1694, doi:10.1175/MWR-D-15-0242.1.

Blaylock, B., 2017. How to initialize WRF with HRRR boundary conditions. URL: http://home.chpc.utah.edu/~u0553130/Brian_Blaylock/hrrr.html.

Blaylock, B., J. Horel, E. Crosman, 2017. Impact of Lake Breezes on Summer Ozone Concentrations in the Salt Lake Valley. *Journal of Applied Meteorology and Climatology*, 56, 353-370, doi: 10.1175/JAMC-D-16-0216.1.

Bowman, D., J. Lees, 2015. Near real time weather and ocean model data access with rNOMADS. *Computers & Geosciences*, 78, 88-95. doi: 10.1016/j.cageo.2015.02.013.

Center for High Performance Computing, 2017. Storage services at CHPC. URL: https://www.chpc.utah.edu/resources/storage_services.php.

Chou, D., 2015. Cloud computing: A value creation model. *Computer Standards & Interfaces*, 38, 72-77. doi: 10.1016/j.csi.2014.10.001.

Climate Prediction Center, 2017. WGRIB2: Utility to read and write grib2 files. URL: http://www.cpc.ncep.noaa.gov/products/wesley/wgrib2/.

Crosman, E., J. Horel, 2017. Large-eddy simulations of a Salt Lake Valley cold-air pool. *Atmospheric Research*, 193, 10-25. doi: 10.1016/j.atmosres.2017.04.010.

Foster, C., E. Crosman, J. Horel, 2017. Simulations of a Cold-Air Pool in Utah's Salt Lake Valley: Sensitivity to Land Use and Snow Cover. *Boundary-Layer Meteorology*, 164, 63-87. doi: 10.1007/s10546-017-0240-7.

Horel, J., and Coauthors, 2002. Mesowest: Cooperative Mesonets in the Western United States. *Bulletin of the American Meteorological Society*, 83, 211–225, doi: 10.1175/1520-0477(2002)083<0211:MCMITW>2.3.CO;2.

Horel, J., E. Crosman, A. Jacques, B. Blaylock, S. Arens, A. Long, J. Sohl, R. Martin, 2016. Influence of the Great Salt Lake on summer air quality over nearby urban areas. *Atmospheric Science Letters*, 17, 480-486. doi: 10.1002/asl.680.

Ikeda, K., M. Steiner, G. Thompson, 2017. Examination of mixed-phase precipitation forecasts from the High-Resolution Rapid Refresh model using surface observations and sounding data. *Weather and Forecasting*, 32, 949-967. doi: 10.1175/WAF-D-16-0171.1.

Jacques, A., J. Horel, E. Crosman, F. Vernon, J. Tytell, 2016. The Earthscope US Transportable Array 1 Hz Surface Pressure Dataset. *Geoscience Data Journal*, 3, 29–36. doi: 10.1002/gdj3.37.

James, E., S. Benjamin, M. Marquis, 2017. A unified high-resolution wind and solar dataset from a rapidly updating numerical weather prediction model. *Renewable Energy*, 102, 390-405. doi: 10.1016/j.renene.2016.10.059.

Katona, B., P. Markowski, C. Alexander, S. Benjamin, 2016. The Influence of Topography on Convective Storm Environments in the Eastern United States as Deduced from the HRRR. *Weather and Forecasting*, 31, 1481-1490. doi: 10.1175/WAF-D-16-0038.1.

Kruger, A., R. Lawrence, E. Dragut, 2006. Building a terabyte NEXRAD radar database for hydrometeorology research. *Computers and Geosciences*, 32, 247-258. doi: 10.1016/j.cageo.2005.06.001.

Li, Y. and coauthors, 2017. A Numerical Study of the June 2013 Flood-Producing Extreme Rainstorm over Southern Alberta. *J. Hydrometeor*, 18, 2057-2078, doi: 10.1175/JHM-D-15-0176.1.

Mell, P., T. Grance, 2011. The NIST Definition of Cloud Computing: Recommendations of the National Institute of Standards and Technology. URL: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf. doi: 10.6028/NIST.SP.800-145.

Maltzahn, C., E. Molina-Estolano, A. Khurana, A. Nelson, S. Brandt, S. Weil, 2010. Ceph as a Scalable Alternative to the Hadoop Distributed File System. *USENIX Magazine*, 35, 38-49. URL: http://static.usenix.org/publications/login/2010-08/openpdfs/maltzahn.pdf.

Molthan, A., J. Case, J. Venner, R. Schroeder, M. Checchi, B. Zavodsky, A. Limaye, R. O'Brien, 2015. Clouds in the cloud: Weather forecasts and applications within cloud computing environments. *Bulletin of the American Meteorological Society*, 96, 1369-1379. doi: 10.1175/BAMS-D-14-00013.1.

Moody D., M. Warren, S. Skillman, R. Chartrand, S. Brumby, R. Keisler, T. Kelton, M. Mathis, 2016. Building a living atlas of the Earth in the cloud. *50th Asilomar Conference on Signals, Systems and Computers*. 1273-1277. doi: 10.1109/ACSSC.2016.7869578.

Nawaz, H., G. Juve, R. da Silva, E. Deelman, 2016. Performance Analysis of an I/O-Intensive Workflow executing on Google Cloud and Amazon Web Services. *Parallel and*

Distributed Processing Symposium Workshops, 2016 IEEE International. IEEE, 2016. 535-544. doi: 10.1109/IPDPSW.2016.90.

Poat, M., J. Lauret, W. Betts, 2015. POSIX and Object Distributed Storage Systems Performance Comparison Studies with Real-Life Scenarios in an Experimental Data Taking Context Leveraging OpenStack Swift & Ceph. *Journal of Physics: Conference Series*. 664, 1-9. doi: 10.1088/1742-6596/664/4/042031.

Powers, J. and coauthors, 2017. The Weather Research and Forecasting (WRF) Model: Overview, System Efforts, and Future Directions. *Bulletin of the American Meteorological Society*, In Press. doi: 10.1175/BAMS-D-15-00308.1.

RedHat 2017: CephFS: Ceph File System. URL: http://docs.ceph.com/docs/master/cephfs/.

Sandholm, T., D. Lee, 2014. Notes on Cloud computing principles. *Journal of Cloud Computing*. 3:21, 1-10. doi: 10.1186/s13677-014-0021-5.

Schroeder, J., W. Burgett, K. Haynie, I. Sonmez, G. Skwira, A. Doggett, J. Lipe, 2005. The West Texas mesonet: a technical overview. *Journal of Atmospheric and Oceanic Technology*, *22*, 211-222. doi: 10.1175/JTECH-1690.1.

Silver, J., C. Zender, 2017. The compression-error trade-off for large gridded data sets. *Geoscientific Model Development*, 10, 413-423. doi: 10.5194/gmd-10-413-2017.

Storm Prediction Center, 2017. Storm Reports for 4-5 April 2017. URL: http://www.spc.noaa.gov/climo/reports/170404_rpts.html and http://www.spc.noaa.gov/climo/reports/170405_rpts.html.

SynopticLabs, 2017: MesoWest & SynopticLabs - Fostering collaboration within the weather observing community. URL: https://synopticlabs.org/.

Taylor, K., R. Stouffer, G. Meehl. 2012: An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, 93, 485-498. doi: 10.1175/BAMS-D-11-00094.1.

Wood, N., 2017: RCLONE- rsync for cloud storage. URL: https://rclone.org/.

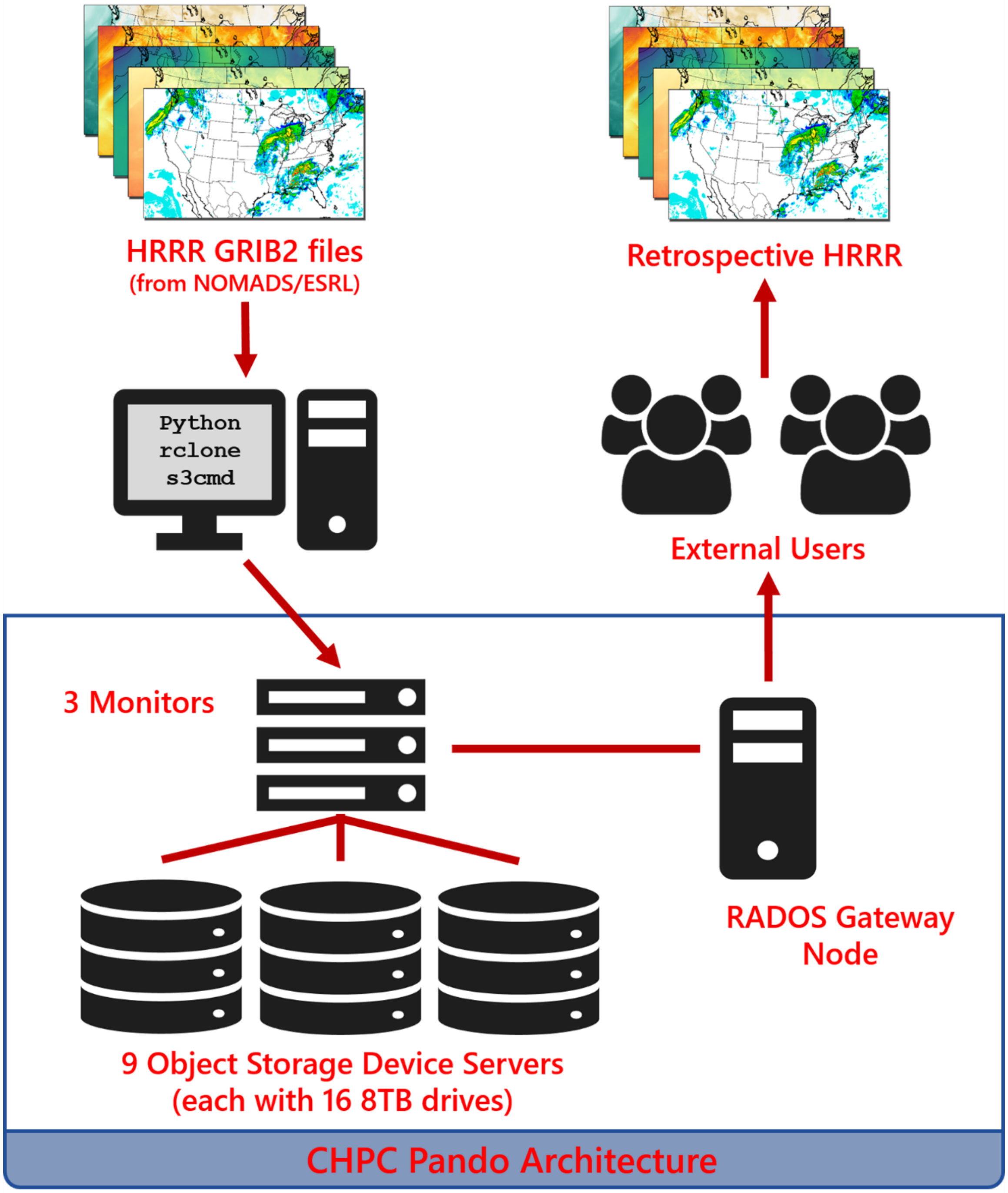


Fig. 1. Present architecture of the Pando archive system.

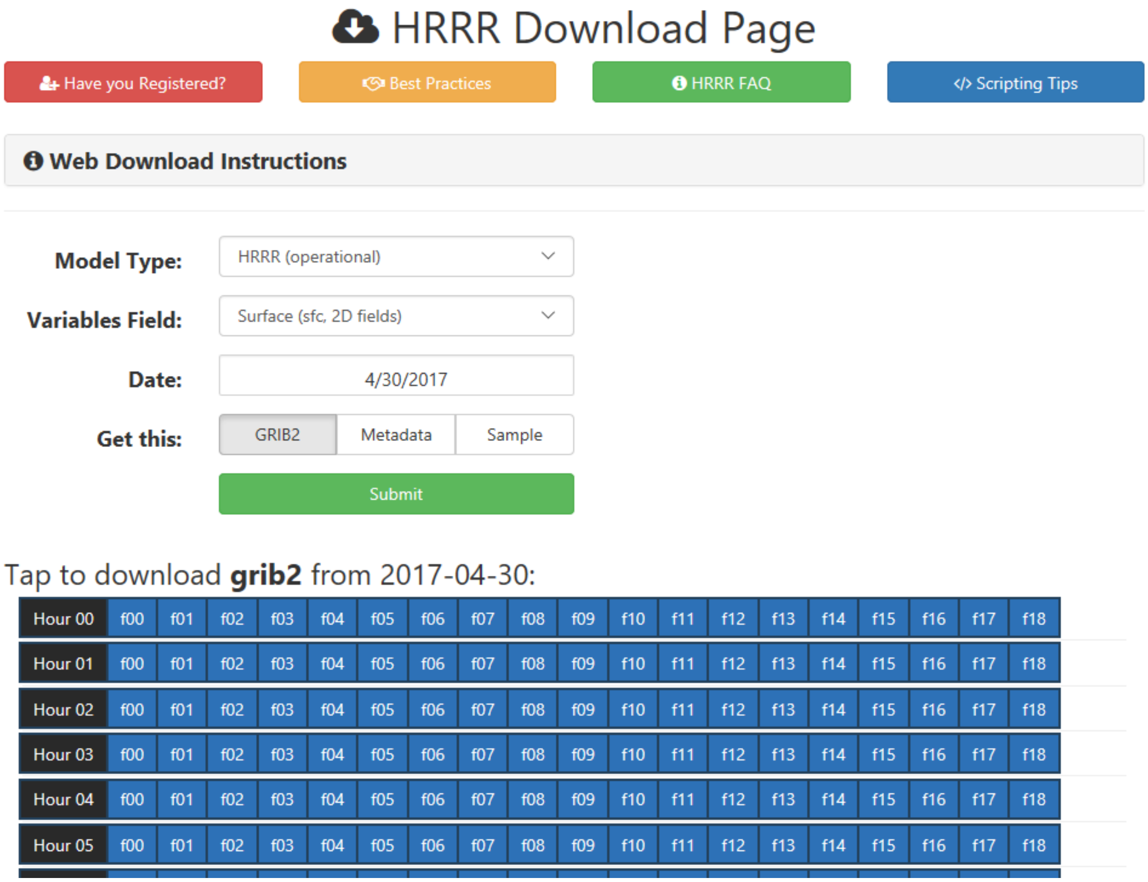


Fig. 2. Web interface to interactively access HRRR model output at http://hrrr.chpc.utah.edu.

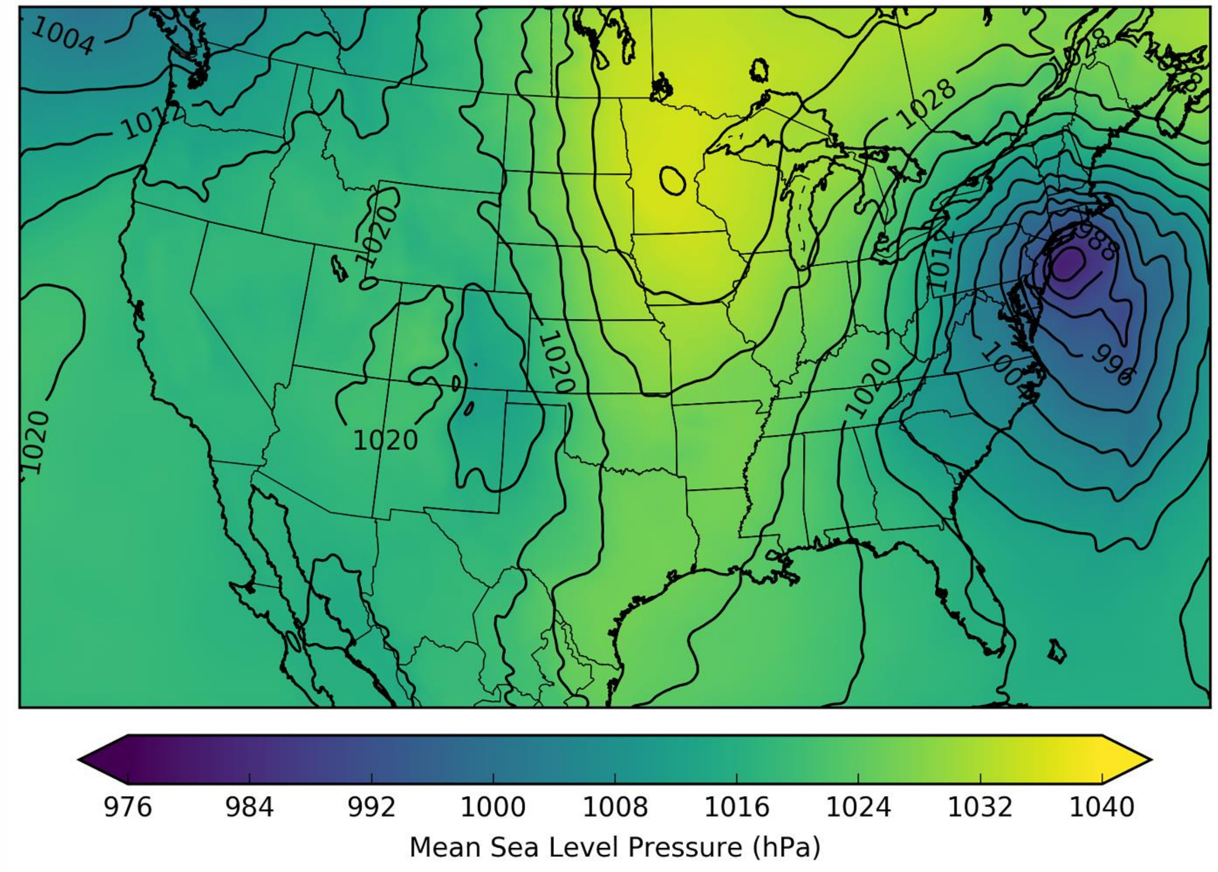


Fig. 3. Mean sea level pressure (hPa) from HRRR analysis at 1700 UTC 14 March 2017 during a high impact New England snowstorm.

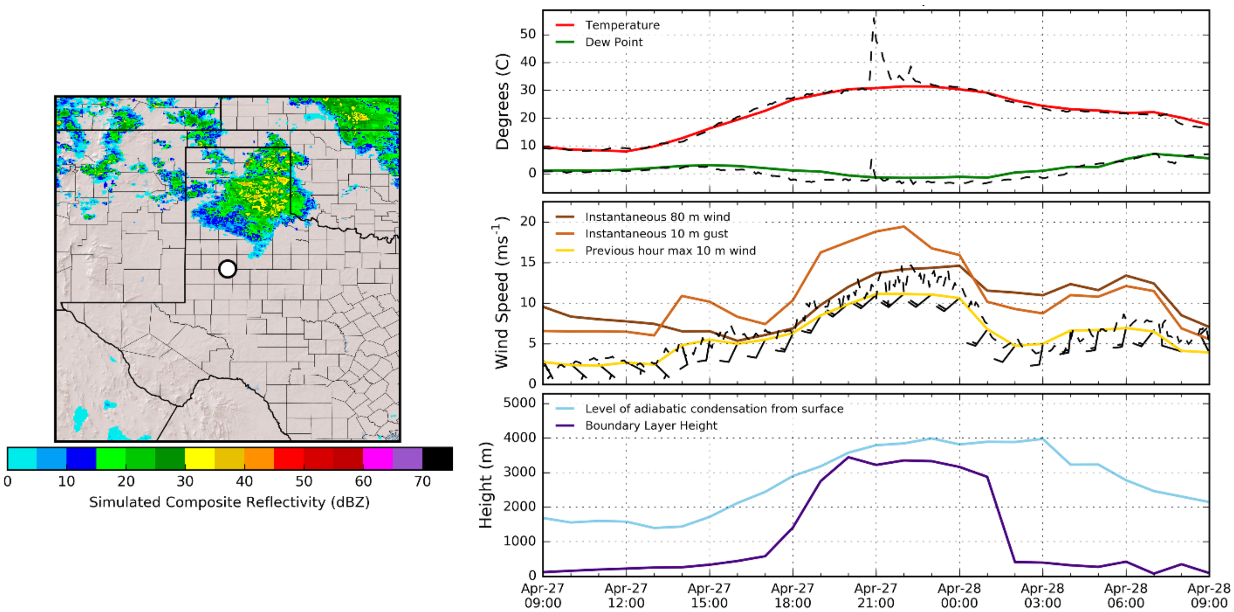


Fig. 4. (Left) HRRR simulated radar reflectivity (dBZ) at 2100 UTC 27 April 2017 at the time of a wildfire near O'Donnell, Texas (white circle). (Right) HRRR analysis of temperature (^{o}C), dew point temperature (^{o}C), 80 m wind speed (m s^{-1}), 10 m gust (m s^{-1}), 10 m maximum wind speed (m s^{-1}), 10 m wind speed and direction (half and full barbs denote 2.5 and 5 m s^{-1}, respectively and direction from which the wind blows denoted by the shaft), boundary layer height (m), and level of adiabatic condensation (m) between 0900 UTC 27 April 2017 and 900 UTC 28 April 2017 near O'Donnell, Texas (white circle on the left). Observed temperature, dew point temperature, and wind speed from the O'Donnell West Texas mesonet site are shown by dashed black lines in the upper two panels.

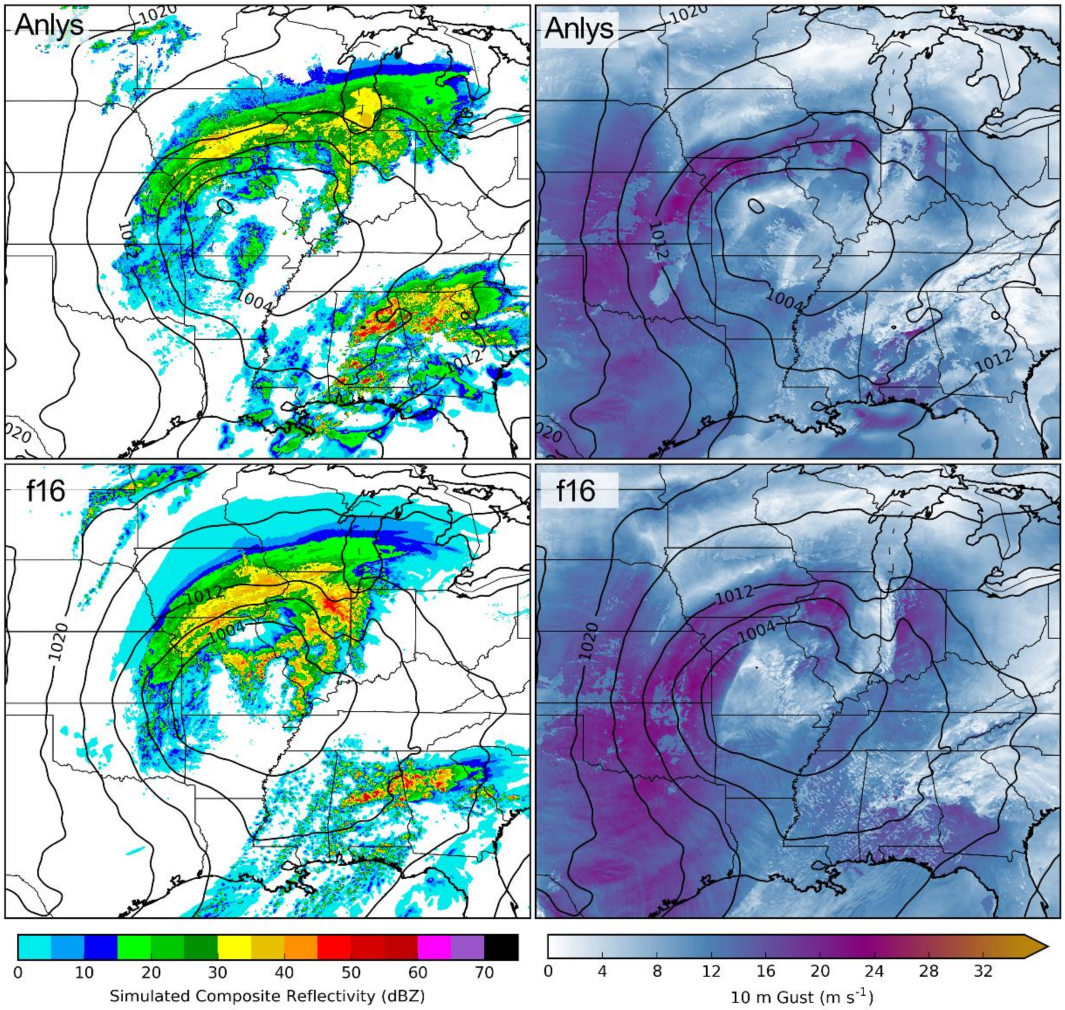


Fig. 5. HRRR analyses (top panels) and HRRR 16 h forecasts (bottom panels) of mean sea level pressure (contours at intervals of 4 hPa) valid 1400 UTC 5 April 2017 with simulated composite radar reflectivity (left panels in dBZ) and 10 m gusts (right panels in m s^{-1}).

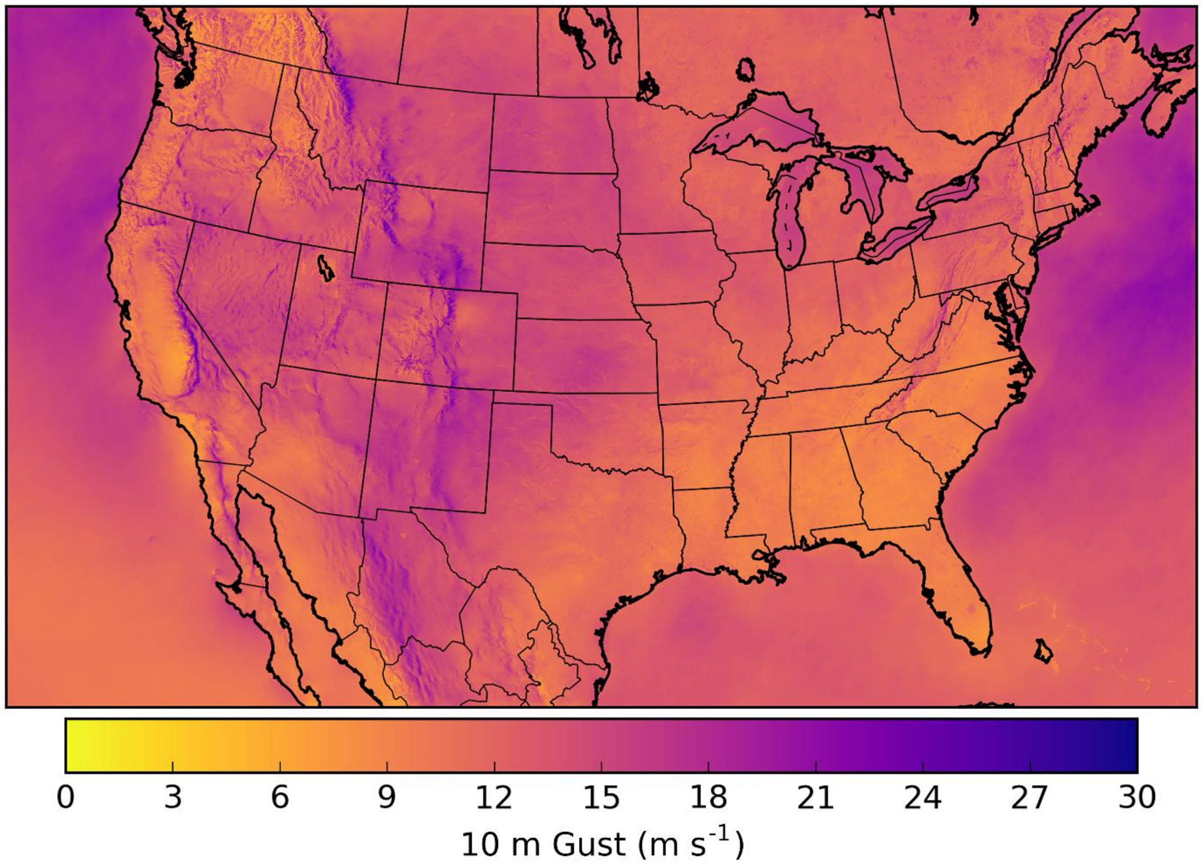


Fig. 6. 95th percentile 10 m gusts (m s^{-1}) from HRRR analyses at 2300 UTC for all days between 18 April 2015 and 30 March 2017.