Big Ensemble Data Assimilation in Numerical Weather Prediction

Takemasa Miyoshi, RIKEN Advanced Institute for Computational Science, University of Maryland, and Japan Agency for Marine-Earth Science and Technology

Keiichi Kondo and Koji Terasaki, RIKEN Advanced Institute for Computational Science

Powerful computers and advanced sensors enable precise simulations of the atmospheric state, requiring data assimilation to connect simulations to real-world sensor data using statistical mathematics and dynamical systems theory. Numerical weather prediction (NWP) thus enables simulations that more closely represent the real world. The authors explore the NWP-associated challenges in managing big data through supercomputing.

High-performance computing (HPC) is essential for *numerical weather prediction* (NWP), the method by which computer models of the atmosphere are used to predict the weather. Advances in computing power enable higher resolution and more complex physical representations of the atmosphere. Although these more advanced representations have led to more accurate weather forecasts from supercomputers than the first models from 1950, the technology is still far from ideal.[1]

In NWP, synchronizing the computer simulation with the real world is essential to accurately determine the atmosphere's current state and likely evolution. Although more precise simulations and more powerful computing are helpful in improving accuracy, data assimilation (DA) plays a key role in improving integration between the computer simulation and real-world observation data.[2,3] DA also employs HPC; in fact, global NWP systems devote equivalent computational resources to DA and 10-day forecast simulation.

To accurately represent the probability density function (PDF) in the ensemble Kalman filter (EnKF)—an advanced DA approach widely used in NWP—within the global atmosphere, we used a large sample size and the

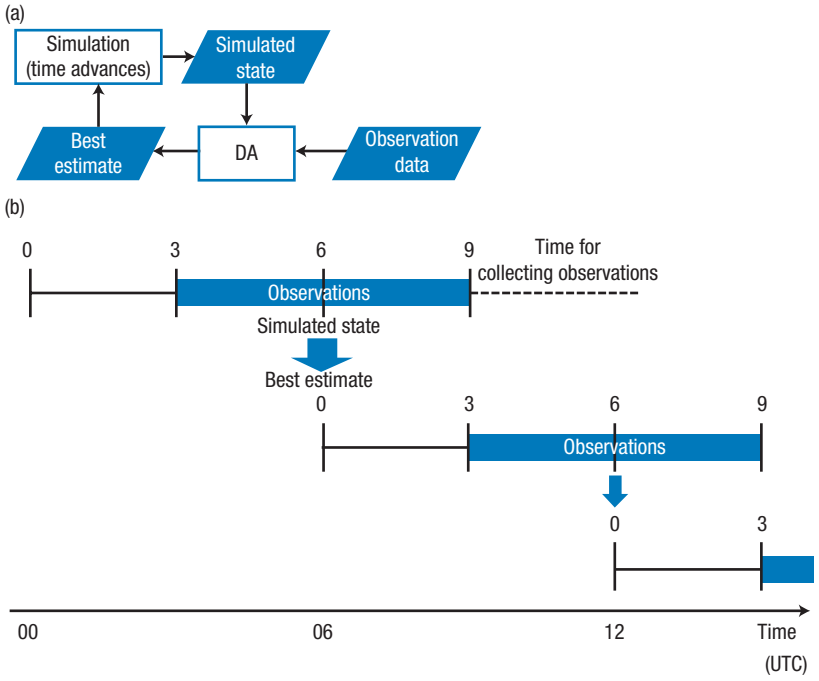

FIGURE 1. Data assimilation (DA) cycles. (a) Flowchart. (b) Timeline for global numerical weather prediction (NWP). The lines labeled 0, 3, 6, and 9 (hours) indicate simulations. The dashed line shows the time for collecting observations. The bottom line indicates time in Coordinated Universal Time (UTC).

K computer, one of the world's largest high-performance computers, located in Japan. To the best of our knowledge, this is the first report on the real-world application of global atmospheric ensemble DA with such a large sample size.

DATA ASSIMILATION

The best atmospheric-state estimate is used as the baseline condition upon which subsequent simulations are based as time advances. DA can yield the most accurate atmospheric state estimate through a computer simulation/real-world observation combination (see Figure 1a). The simulation DA cycle is then repeated and new observations are incorporated. For global NWP, the cycle interval is six hours (see Figure 1b). This is usually done using a nine-hour simulation that we then make simulated-state corrections to at hour six using the observations from hours three to nine (see Figure 1b). In real-time applications, we wait several hours before collecting observations (the dashed line in Figure 1b), so that the best estimate at hour six is executed at about 12:00 Coordinated Universal Time (UTC) or later. The exact wait time for collecting observations varies among different NWP centers. This way, the simulation DA cycle is executed retrospectively after a delay of several hours. The best estimate at hour six initializes the nine-hour simulation for the next DA cycle, and also initializes a 10-day forecast simulation. The exact forecast length depends on the NWP centers; for example, the US National Centers for Environmental Prediction (NCEP) run 16-day forecasts, and the Japan Meteorological Agency (JMA) runs 11-day forecasts.

DA finds the best mix of the simulated state and observations based on Bayes' rule. Here, the simulated-state PDF is multiplied by the observation likelihood function, which is normalized to obtain the final PDF. The maximizer of this PDF is considered the best estimate—the maximum likelihood estimator.

Because the degrees of freedom of the atmospheric state are large, typically $O(10^8)$ or more, it is prohibitive to explicitly represent the PDF. Moreover, in theory, the simulated state's PDF is obtained through an evolution of the PDF from the previous DA step's best estimate, but calculating the high-dimension PDF evolution is even more prohibitive: a single-state simulation requires HPC, thus a PDF is extremely computationally intensive. A number of simplifying assumptions have therefore been developed and are at the center of DA research; for example, the Gaussian assumption and a limited-sample Monte Carlo ensemble representation of the PDF enable the EnKF.[4–6] The sample size for EnKF is usually limited to about 100 samples, because each requires its own expensive simulation. Although running 100 parallel simulations requires significant computational resources, the sample size is nevertheless too limited to accurately represent the high-dimensional PDF of the atmospheric state. Most EnKF studies have focused on using these very limited samples more wisely to improve the computational speed and analytical accuracy of EnKF.

We took a different approach by using an extremely large sample size. The sample size and the simulation complexity are both important, but we ran a large sample size at the expense of resolution to explore the potential benefits of doing so.

SIMULATED OBSERVATIONS WITH AGCM

A previous EnKF experiment[7] used 10,240 samples (the largest known sample size for the global atmosphere) and an intermediate *atmospheric general circulation model* (AGCM) known as SPEEDY,[8] and investigated the spatial correlation patterns and Gaussian

TABLE 1. Data sizes and computational costs for SPEEDY.

Sample size	Observation data size (Kbytes)	Simulation data size	Computational cost (ensemble forecasts; wall-clock time [s])	Computational cost (ensemble forecasts; node no.)	Computational cost (ensemble forecasts; node hour)	Computational cost (LETKF*; wall-clock time [s])	Computational cost (LETKF*; node no.)	Computational cost (LETKF*; node hour)
80	346	44.2 Mbytes	3.0	80	0.067	4.7	80	0.1
10,240	346	5.66 Gbytes	19.5	4,608	25.000	285.0	4,608	365.0

*LETKF: local ensemble transform Kalman filter

nature of the PDF. The SPEEDY model is a typical AGCM with five prognostic variables: horizontal wind components (U, V); temperature (T); specific humidity (Q); and surface pressure (P_s). These variables are defined at 96 by 48 (total 4,608) horizontal grid points in the longitudinal and latitudinal directions, respectively. The first four variables—U, V, T, and Q—expand in the 3D space with seven vertical grid points ranging from the lower troposphere (the lowest layer of Earth's atmosphere; about 925 hectopascal [hPa]) to the tropopause (the boundary between the troposphere and the stratosphere; about 100 hPa).

Fast computation is a major advantage of using the SPEEDY model; it takes less than one second to run a single six-hour simulation using a single CPU core. For DA, we use the local ensemble transform Kalman filter (LETKF),[9] which is particularly efficient for parallel architecture because LETKF independently solves the EnKF equations at each grid point. The LETKF equations include an eigenvalue decomposition of an m-by-m symmetric matrix, where m is the sample size. When m is large, the LETKF computations become expensive, as typical eigenvalue solvers require $O(m^3)$ computations. Even with the efficient eigenvalue solver known as EigenExa[10] (www.aics.riken.jp/labs/lpnctrt/EigenExa_e.html), it takes about 4.75 minutes with 4,608 nodes of the K computer (peak ~580 teraflops [Tflops]) for a single DA step with m = 10,240 samples. Here, we assumed typical upper-air observations at 416 stations, totaling 10,816 observations per DA step. Table 1 summarizes the data sizes and computational costs for SPEEDY.

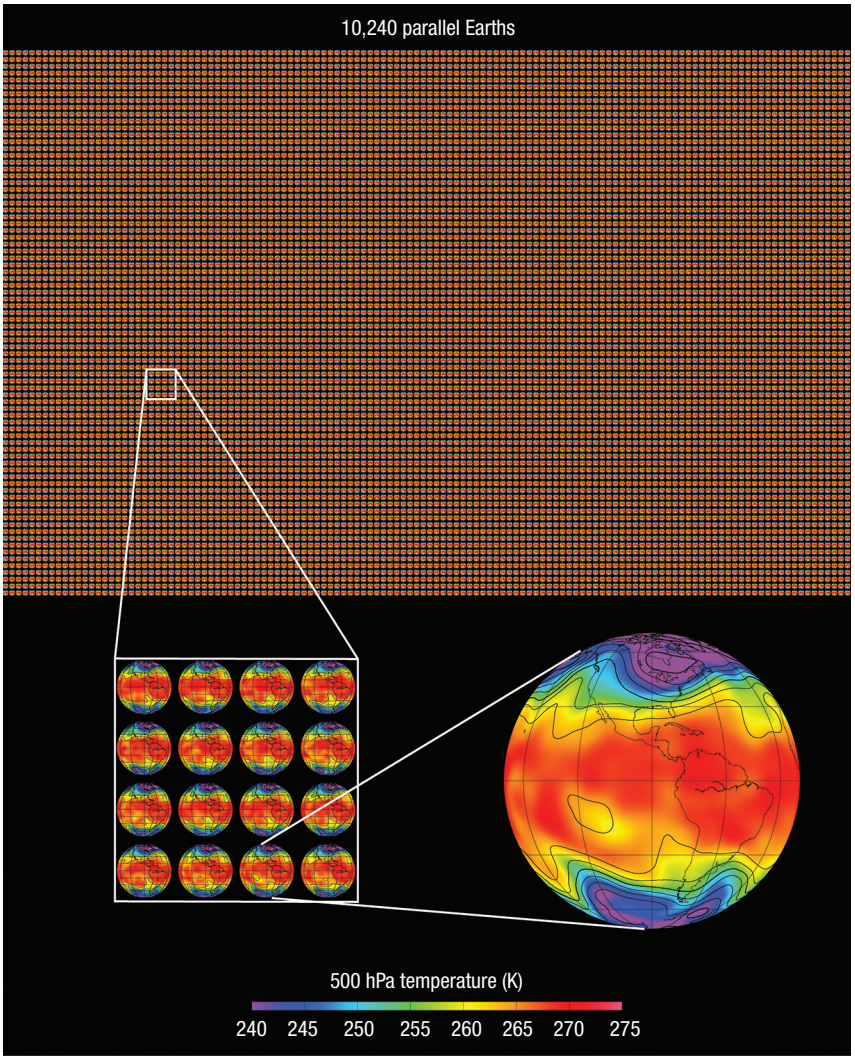

FIGURE 2. Atmospheric sample states after the first DA step. Sixteen of the total 10,240 samples are magnified to show the atmospheric states, and 1 out of these 16 is further magnified. hPa is hectopascal.

Figure 2 illustrates atmospheric states' large sample size. Each small circle corresponds to the complete atmosphere, and 10,240 parallel Monte Carlo simulations were performed to assimilate observations every six hours. Because the circles are small, 16 of the total samples are magnified to show the atmospheric states,

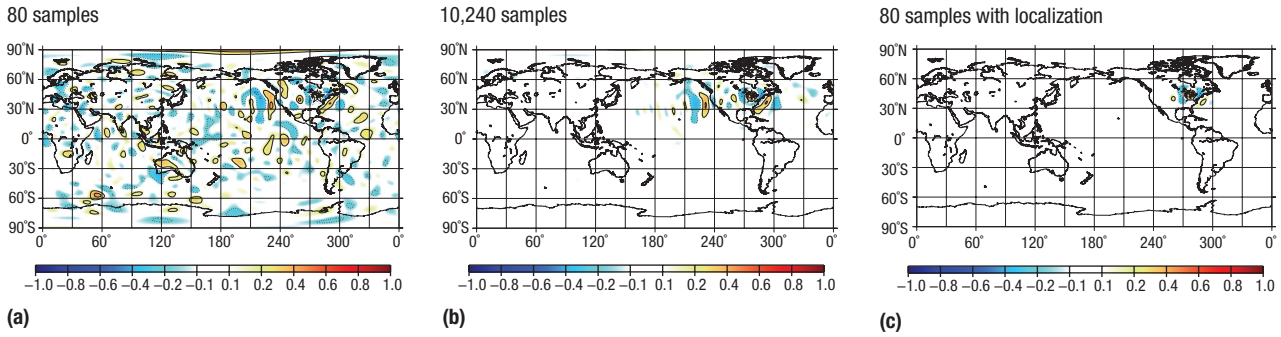

FIGURE 3. Horizontal maps of the autocorrelations for specific humidity around the 500-hPa level from the star (42.678 N, 82.50 W) using (a) 80 samples and (b) 10,240 samples; (c) shows the localization of (a).

TABLE 2. Data sizes and computational costs for the Nonhydrostatic Icosahedral Atmospheric Model.

Sample size	Observation data size (Kbytes)	Simulation data size	Computational cost (ensemble forecasts; wall-clock time [s])	Computational cost (ensemble forecasts; node no.)	Computational cost (ensemble forecasts; node hour)	Computational cost (LETKF; wall-clock time [s])	Computational cost (LETKF; node no.)	Computational cost (LETKF; node hour)
80	960	9.47 Gbytes	726	200	41	48	200	2.7
10,240	960	1.21 Tbytes	3,666	6,400	6,500	1,310	5,784	2,100.0

and 1 of these 16 is further magnified. As this was a simulation study, no real data was used. Denser observations were simulated over the populated areas, where more similarities were found among samples when looking closely at the magnified atmospheric states. All the simulations represent equally probable estimates of the atmospheric state. The large sample size reduced the sampling error and significantly improved the accuracy of DA and forecasts.

Following the previous EnKF study,[7] we focused on the direct impact of the large sample size by looking at horizontal maps of the autocorrelations for humidity with 10,240 samples and 80 subsamples, which were randomly chosen from the total samples. Figure 3 illustrates how DA works, showing the impact of an observation at the yellow star. With 80 samples (shown in Figure 3a), the impact of the observation around the Great Lakes in the US spreads to the entire globe. This is unrealistic and is likely due to sampling error. Thus, we localized the impact of observations within the adjacent area (shown in Figure 3c). The localization radius was optimized manually by trial and error, and we used the Gaussian function with a 1,400-kilometer (km) standard deviation to optimize the DA accuracy. With 10,240 samples (shown in Figure 3b), the sampling error was greatly reduced, and the significant correlations spread toward the eastern Pacific beyond several thousand kilometers.

REAL-WORLD EXPERIMENT

We recently performed another EnKF experiment with 10,240 samples for the real global atmosphere, using the LETKF system with the Nonhydrostatic Icosahedral Atmospheric Model (NICAM)[11,12] or NICAM-LETKF.[13] Unlike the SPEEDY model, NICAM is designed to precisely simulate the actual atmosphere with state-of-the-art components of physics modules such as radiation, sub-grid-scale turbulences, and water-related physics. We employed relatively low horizontal resolution at approximately 112 km, sufficient for representing typical mid-latitude weather systems; the 40 vertical levels extend up to 40 km (about 3 hPa). NICAM's prognostic variables include those of SPEEDY (U, V, T, Q, and P_s) as well as vertical wind (W), 3D pressure (P), and cloud liquid water content (CLW).

NICAM's grid is based on the 20-triangle icosahedron and nearly covers the globe homogeneously. Each triangle is divided into four to double the resolution, and the division is repeated until the desired resolution is reached. For the current 112-km resolution setting, the division is repeated six times (grid division level 6), and the number of grid points is 40,960. The global grid (118 Mbytes) is separated into 40 rhombuses, and each rhombus region has its own file I/O (~3 Mbytes). Thus, a single snapshot of the NICAM atmospheric state consists of 40 files. With 10,240 samples, the number of files reaches 409,600. For

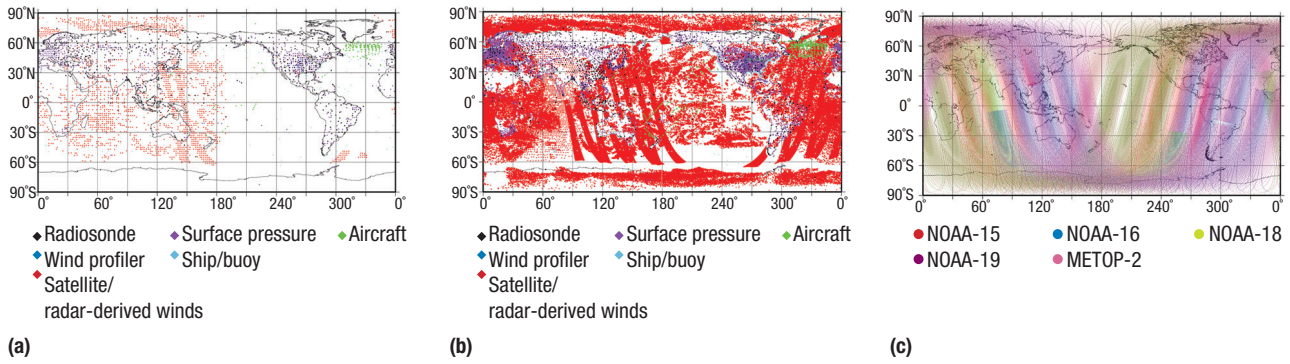

FIGURE 4. Observation coverage for DA. (a) Reduced set of the 26,268 observations input to DA at 12:00 UTC on 1 July 2011. Each color corresponds to different observing platforms. (b) Raw PREPBUFR data without thinning, totaling 553,471 observations. (c) Satellite-borne Advanced Microwave Sounding Unit-A data from five spacecraft (NOAA-15, -16, -18, and -19, and METOP-2); approximately 81,000 scanning locations for each spacecraft.

the input to DA, we usually split the six-hour DA window into seven hourly slots, and the number of files reaches nearly three million at every DA step. Handling this large number of files could be a challenge.

The real-world experiment followed previous experiments with real observations from NCEP known as PREPBUFR,[13] which include reports from weather balloons (radiosondes), wind profilers, aircrafts, ships, buoys, and surface stations, as well as wind data derived from satellites and radar (see Figure 4). Although studies have suggested that the satellite-borne Advanced Microwave Sounding Unit (AMSU)-A should be the highest-impact data for global NWP,[14] we did not use these here.

The raw PREPBUFR data includes dense observations (see Figure 4b), and we reduced the data spatially and temporally according to the NICAM resolution (112 km) and to save memory space for large ensemble DA. This way, we used 2.6×10^4 observations (1 Mbyte; see Figure 4a), about 5 percent of the total 5.5×10^5 observations (20 Mbytes). Using 5,784 nodes of the K computer (peak ~720 Tflops), a single DA step takes about 21.8 minutes with m = 10,240 samples. Table 2 summarizes the data sizes and computational costs for NICAM.

We ran NICAM-LETKF with 10,240 samples for a week, from 00:00 UTC on 1 November 2011 to 00:00 UTC on 8 November 2011. The simulated states quickly converged to the real atmosphere, and the large sample size shows general improvements to DA accuracy. Similar to the simulated observations, we focused on the horizontal autocorrelation patterns (see Figure 5). With 80 samples (see Figure 5a), the correlation patterns were spread all over, so it was difficult to distinguish the signal from the noise. By applying manually optimized localization, we removed the long-range correlations regardless of whether the signal was present (see Figure 5c). With more samples, we can reduce the sampling noise and find the significant correlation patterns (see Figure 5b).

As shown in Figure 5, the filament-like correlations from the star over the Great Lakes reach well beyond the North American coasts: eastward to Scandinavia and westward to the tropical Pacific. An observation over the Great Lakes could potentially contain information about the atmospheric state along the filaments over the Pacific and Atlantic. We found that the filament-like patterns correspond well to the atmospheric flow and would likely change from day to day and place to place. Our quick investigation revealed that such planetary-scale correlation patterns certainly exist, but only occasionally. In many cases, the horizontal correlations are much more localized within a continent.

FUTURE CHALLENGES

The results from our large ensemble DA experiment could provide important information for developing efficient EnKF methods with smaller ensembles. Further investigations into when and where to find the long-range correlations and how they can be used to improve DA and NWP are important subjects of future meteorological research.

A large number of observations come from satellite remote sensing. Satellite passive sensors observe radiative power emitted from the Earth at various frequency bands. The radiative power is the result of complex radiation processes including surface emission and atmospheric absorption, transmission, scattering, and emission, providing information about atmospheric profiles for temperature, moisture, and various atmospheric constituents such as dust, pollutants, and carbon dioxide. Satellite data provides critical information to contemporary NWP systems. The highest-impact AMSU-A mainly provides information about the vertical profile of atmospheric temperature. AMSU-A data covers the globe almost entirely in six hours from five polar-orbiting satellites (see Figure 4c), and each satellite scans approximately

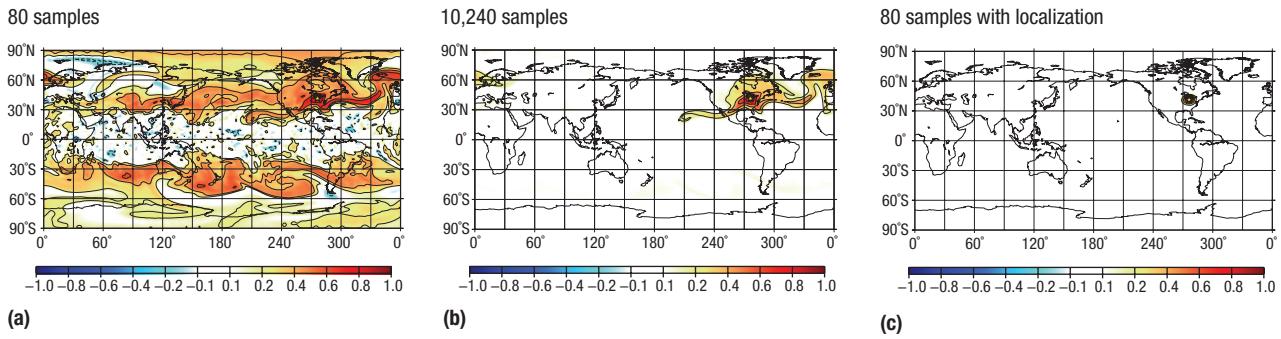

FIGURE 5. Horizontal maps of real atmosphere autocorrelations using NICAM-LETKF. The yellow star has slightly shifted from its placement in Figure 3 to (41.39 N, 83.84 W), and the vertical level is about 100 hPa.

8.1×10^4 locations in six hours. Multiple frequency bands or channels observe each scanning location. AMSU-A has 15 channels, so 6.1×10^6 (58.3 Mbytes) data is obtained every six hours. This is already about 10 times the size of the PREPBUFR data, but we have other satellite sensors including hyperspectral infrared sounders, which have O(1,000) channels for similar scanning locations. Roughly speaking, the data number reaches $O(10^8)$ (1 Gbyte) or more every six hours from multiple satellites. A fraction of this existing data—typically less than a percent—has been used in operational NWP systems because the resolution of DA is usually lower than the satellite scanning resolution, and because so many radiative transfer computations are required. With consistently increasing computer power, the underused satellite data should be included for DA. Sensor technologies also keep advancing; a recent example is the geostationary satellite Himawari-8, which started full operations on 7 July 2015. This satellite scans a side of the globe (the full disk) every 10 minutes, which is three times faster than the previous Himawari-7 (or MTSAT-2), and has higher resolution with more channels, producing about 50 times more data per unit time.

We are now in the big data era for NWP, with orders of magnitude more samples, more observations, and higher resolutions. Dense and frequent observation data could be used more effectively with high-resolution DA. As computers keep advancing, the resolution of DA keeps improving: Yoshiyaki Miyamoto and his colleagues performed a global atmospheric simulation using NICAM at 870-meter resolution—the highest yet—using the K computer.[15] We could call this high-end simulation "big simulation." As computer performance increases, we will most likely integrate big observation data into the big simulation, which produces even bigger data than advanced observations. In the large ensemble SPEEDY experiment, the simulation data size was 5.66 Gbytes at each DA step. For the NICAM experiment at relatively low 112-km resolution, the simulation data size increased to 1.21 Tbytes. As the horizontal resolution doubles, the data size quadruples.

Because we have strict time limits for real-time NWP, handling data and I/O in a timely manner will be challenging. To tackle this, it is important to develop expertise from NWP modeling, DA, observation technology, and computer science. Japan's FLAGSHIP 2020 Project will develop a post-petascale supercomputer—the post-K computer—using codesign, or collaborative development of the system design and applications. The NICAM-LETKF model was chosen as a representative codesign application. The project explores future high-resolution NWP feasibility to take full advantage of big observation data. We believe that collaborating among fields is the key to big data success.

ACKNOWLEDGMENTS

This study was partly supported by CREST, the Japan Science and Technology Agency, the Japan Aerospace Exploration Agency (JAXA) Precipitation Measuring Mission (PMM), and Japan's FLAGSHIP 2020 Project.

REFERENCES

1. J.G. Charney, R. Fjortoft, and J. von Neumann, "Numerical Integration of the Barotropic Vorticity Equation," *Tellus*, vol. 2, no. 4, 1950, pp. 237–254.
2. E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge Univ. Press, 2002.
3. T. Tsuyuki and T. Miyoshi, "Recent Progress of Data Assimilation Methods in Meteorology," *J. Meteorological Society of Japan*, vol. 85B, 2007, pp. 331–361.
4. G. Evensen, "Sequential Data Assimilation with a Nonlinear Quasi-geostrophic Model using Monte Carlo Methods to Forecast Error Statistics," *J. Geophysical Research*, vol. 99, no. C5, 1994, pp. 10143–10162.
5. G. Evensen, "The Ensemble Kalman Filter: Theoretical Formulation and Practical Implementation," *Ocean Dynamics*, vol. 53, no. 4, 2003, pp. 343–367.
6. P.L. Houtekamer and H.L. Mitchell, "Data Assimilation using an Ensemble Kalman Filter Technique,"

Monthly Weather Review, vol. 126, no. 3, 1998, pp. 796–811.
7. T. Miyoshi, K. Kondo, and T. Imamura, "The 10,240-Member Ensemble Kalman Filtering with an Intermediate AGCM," *Geophysical Research Letters*, vol. 41, no. 14, 2014, pp. 5264–5271.
8. F. Molteni, "Atmospheric Simulations using a GCM with Simplified Physical Parametrizations. I: Model Climatology and Variability in Multi-decadal Experiments," *Climate Dynamics*, vol. 20, no. 2, 2003, pp. 175–191.
9. B.R. Hunt, E.J. Kostelich, and I. Szunyogh, "Efficient Data Assimilation for Spatiotemporal Chaos: A Local Ensemble Transform Kalman Filter," *Physica D*, vol. 230, 2007, pp. 112–126.
10. T. Imamura, S. Yamada, and M. Machida, "Development of a High-Performance Eigensolver on a Peta-Scale Next-Generation Supercomputer System," *Progress in Nuclear Science and Technology*, vol. 2, 2011, pp. 643–650.
11. H. Tomita and M. Satoh, "A New Dynamical Framework of Nonhydrostatic Global Model Using the Icosahedral Grid," *Fluid Dynamics Research*, vol. 34, no. 6, 2004, pp. 357–400.
12. M. Satoh et al., "The Non-hydrostatic Icosahedral Atmospheric Model: Description and Development," *Progress in Earth and Planetary Science*, vol. 1, no. 18, 2014; doi: 10.1186/s40645-014-0018-1.
13. K. Terasaki, M. Sawada, and T. Miyoshi, "Local Ensemble Transform Kalman Filter Experiments with the Nonhydrostatic Icosahedral Atmospheric Model NICAM," *SOLA*, vol. 11, 2015, pp. 23–26.
14. Y. Ota et al., "Ensemble-based Observation Impact Estimates Using the NCEP GFS," *Tellus*, vol. 65A, 2013; www.tellusa.net/index.php/tellusa/article/view/20038.
15. Y. Miyamoto et al., "Deep Moist Atmospheric Convection in a Subkilometer Global Simulation," *Geophysical Research Letters*, vol. 40, no. 8, 2013, pp. 4922–4926.

ABOUT THE AUTHORS

TAKEMASA MIYOSHI leads the Data Assimilation Research Team at RIKEN Advanced Institute for Computational Science, is a visiting professor in the Department of Atmospheric and Oceanic Science at the University of Maryland, and is a visiting senior scientist in the Application Laboratory at the Japan Agency for Marine-Earth Science and Technology. His research interests include numerical weather prediction, data assimilation theory, and applications. Miyoshi received a PhD in meteorology from the University of Maryland. He is a member of the American Meteorological Society, the American Geophysical Union, the Meteorological Society of Japan, and the Japan Geoscience Union. Contact him at takemasa.miyoshi@riken.jp.

KEIICHI KONDO is a postdoctoral researcher at RIKEN Advanced Institute for Computational Science. His research interests include data assimilation, numerical weather prediction, and high-performance computing. Kondo received a PhD in science from the University of Tsukuba. He is a member of the Meteorological Society of Japan. Contact him at keiichi.kondo@riken.jp.

KOJI TERASAKI is a research scientist at RIKEN Advanced Institute for Computational Science. His research interests include data assimilation and atmospheric general circulation. Terasaki received a PhD in science from the University of Tsukuba. He is a member of the Meteorological Society of Japan. Contact him at koji.terasaki@riken.jp.

See **www.computer.org/computer-multimedia** for multimedia content related to this article.