The Use of a Portable Parallel Data Processing and Error Analysis System (DPEAS) for Technology Transition of Complex Multi-Satellite Data Fusion Algorithms into Operations

Dr. Andrew S. Jones

AMS Annual Meeting

Session: Research to Operations Pathway for Satellite Data Retrieval Algorithms – Part 1

January 10, 2013

Thanks to our Co-Authors

- Stan Kidder (CSU/CIRA)
- John Forsythe (CSU/CIRA)
- Jiang Zhao (SGT/OSPO)
- Pete Keehn (IMSG/STAR)
- Clay Davenport (SGT/OSPO)
- Limin Zhao (NOAA/NESDIS/OSPO)
- and numerous collaborators…

What is DPEAS?

- DPEAS is a data processing environment
- Developed at CSU under DoD/NOAA funding with ownership by NOAA
  - Runs on Windows/Linux/AIX (natively for each OS)
  - **AIX version is for NOAA/NESDIS/OSPO OPS use**
  - Linux version was used for porting to AIX OPS
  - Built-in automatic parallelism capabilities
- Essentially DPEAS is a modeling/error analysis environment for satellite/model data processing

What is the DPEAS Data Approach?

- Employs a data fabric design using complex multi-satellite algorithm stencils that use shared support libraries
- The approach yields efficiencies and cost savings
  - ADL (JPSS) and AWG (GOES-R) also use similar concepts
- DPEAS grants users full source code access with no predefined configuration limitations since the codes are native for each operating system
- Encourages “clean” algorithm data pathways into NOAA Operations

What is DPEAS?
Some Technical Aspects

- DPEAS is a program that self-replicates using file I/O and remote daemons/services
- Replication is controlled by DPEAS' parallelism constructs and adaptive schedulers
- Typical use is on small machine clusters (~100 logical cores per algorithm job, but can run with 2000+ cores)
- A "Fortran Interpreter" controls high level DPEAS behaviors, which results in low training requirements, stability, and full granular control of the system processes and data fabric components

What is DPEAS?
Modular Support

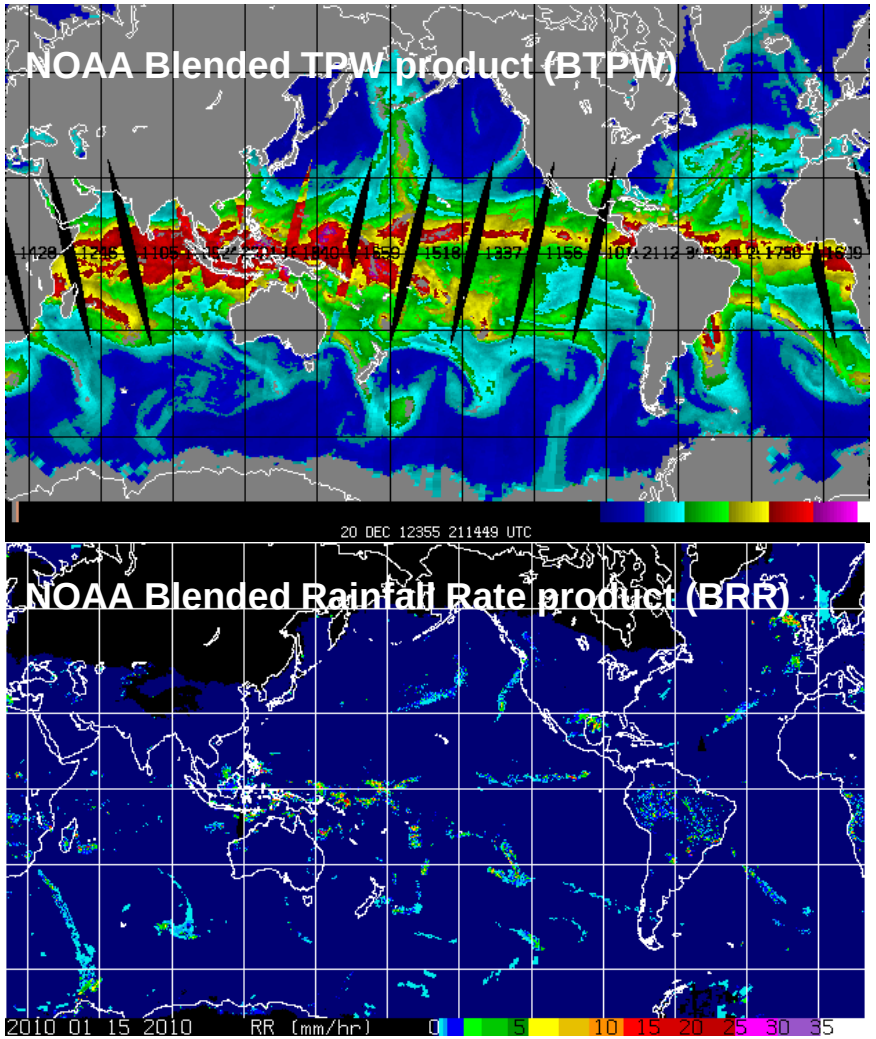
- DPEAS borrows from HDF-EOS for its data fabric structure design – however the data does not reside on disk but within the fabric, providing improved control (Jones and Vonder Haar, 2002)
- DPEAS supports GRIB, McIDAS, binary read/writes, HDFEOS2, HDF4, HDF5, netCDF3, netCDF4, and many other data formats
- Approximately 20+ satellite sensors and several model output sources can be handled by the system
- **Two operational NOAA DPEAS products are in production:**
  1) NOAA Blended Total Precipitable Water (BTPW)
  2) NOAA Blended Rainfall Rate (BRR)
     (contact: Ms. Limin Zhao)
- Updates occur with new sensors about every 6-12 months

DPEAS OPS Examples

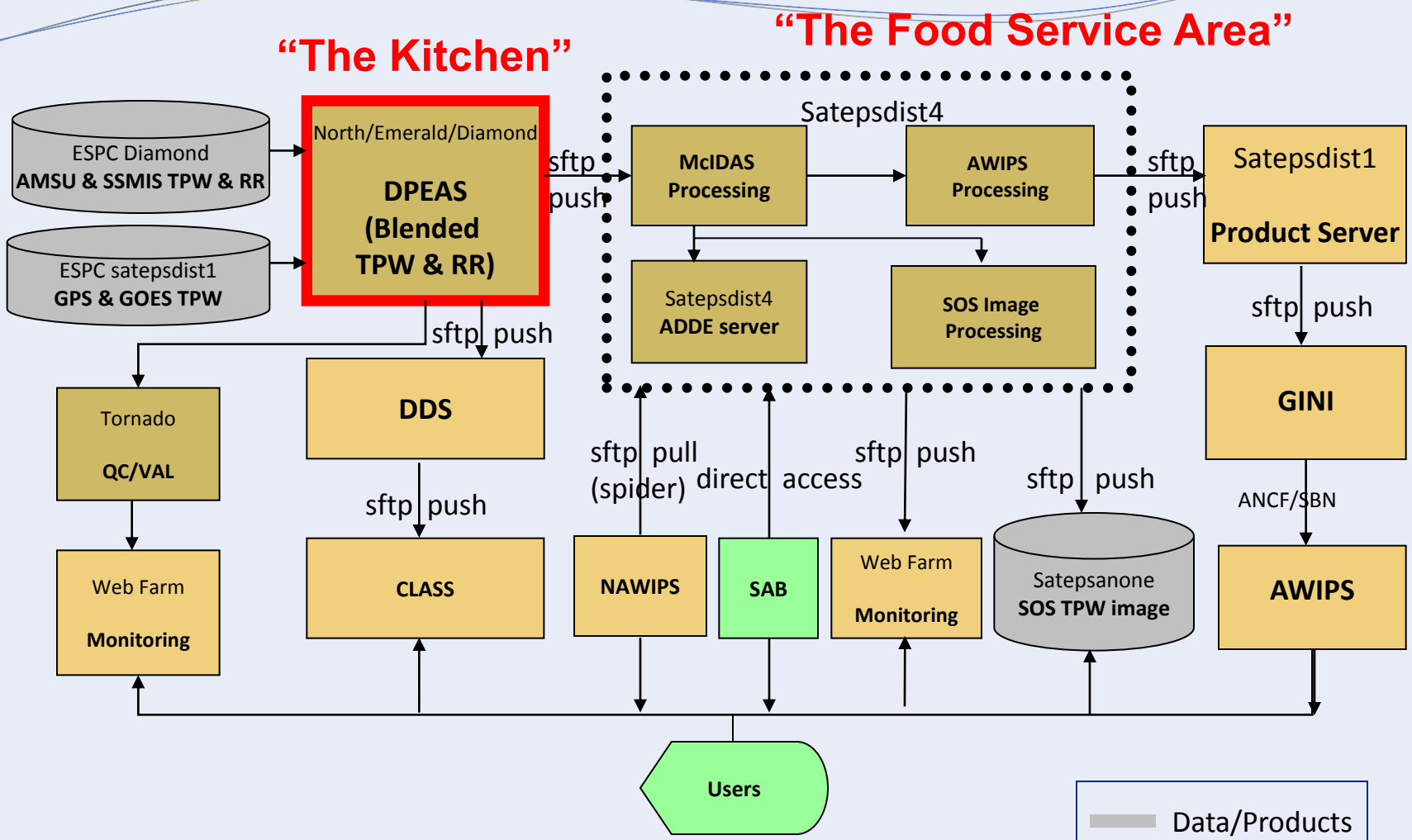
Both products perform an on-the-fly dynamic recalibration to a reference sensor baseline (Kidder and Jones, 2007)

The BRR method is a “cousin” of the BTPW blending algorithm

DPEAS consolidates many complex functions into a single analysis package



The Blended TPW & RR Products System Diagram



Input Data Includes: AMSU TPW & RR from MSPPS, SSMIS TPW from MIRS, SSMIS RR from FNMOC, GOES TPW from SFOV, and GPS from NWS/NOAAPort

A Brief History of DPEAS OPS use

- **1998** – NOAA AMSU development toward blended TPW using the NOAA single sensor products
- **2002** – DPEAS data structures/fabric used as base for NASA CloudSat Data Processing Center (DPC) development activities at CIRA
  - By 2013: 1+ PB of data has been processed and delivered by the CloudSat DPC
- **2004** – Full transition to NESDIS Grid "Experimental" System for OPS
- **2009** – BTPW on AIX declared Operational at NOAA
- **2012** – BRR on AIX declared Operational at NOAA
- Incremental 6-12 month update cycles for insertion of additional sensor results into the BTPW and BRR products
- ATMS OPS, and METOP-B OPS are updates in progress
- **Next: more international sensors**
  JAXA GCOM-W AMSR-2 is targeted for FY13 developments

The Best Part

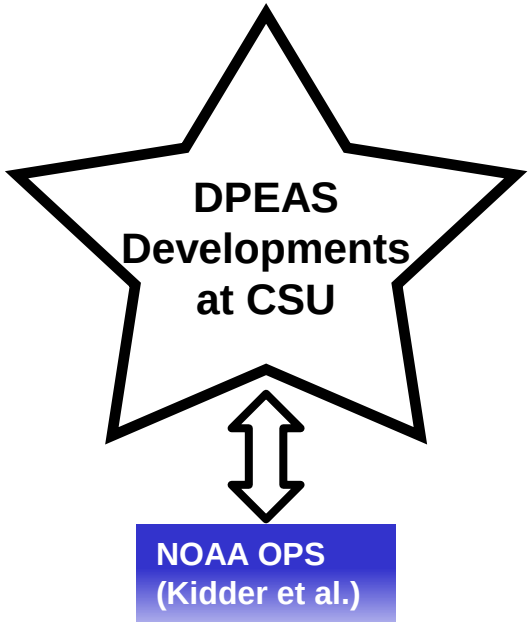
- Algorithm transitions are on the order of days (if the OPS environment remains stable)
- **Some NOAA OPS sensor additions did not require *any* CIRA assistance,** but were achieved through cloning adaptable DPEAS codes (e.g., The NOAA DMSP F-18 SSMIS updates)
- Scheduled project transitions mostly consist of monitoring the transitioned product for anomalies, and keeping up with NOAA requirement updates

Lessons Learned

1. **System code complexity must be minimized** and suitable for a non-specialist to face the reality of contractor turn-over. Elaborate systems tend to drive up costs to avoid possible failure. **"Risk" insurance costs are higher on expensive systems.**
2. **"Don't chase the technology – ride the waves…"** e.g., multi-core machines, evolving "standards" such as netCDF4, cloud computing for redistribution of data.
3. **Small and "partial" solutions are sometimes better than going "large" and perhaps never getting there.**
4. **It's the Science Algorithms vs. the Science Support System**
   We need two mode thinking: (1) fast science change + (2) a stable and slowly-changing support system. The DPEAS Fortran Interpreter and simple parallelism constructs enables the system to change slowly and only when *management* requires it, but DPEAS can also add new science code quickly.

Research to Operations: turned on its head R2O2R (+ Operations to Research)

Ongoing OPS to Research Pathway Examples using DPEAS at CSU:



Research to Operations: turned on its head R2O2R (+ Operations to Research)

Ongoing OPS to Research Pathway Examples using DPEAS at CSU:

CSU Infectious Disease Research (Jones and Corbett)
Use of precipitation data and cloud computing databases

NSF (Fletcher and Jones)
1DVAR Non-Gaussian Data Assimilation

NASA SPoRT (Kidder et al.)
Use of NASA AIRS data for multilayered BTPW

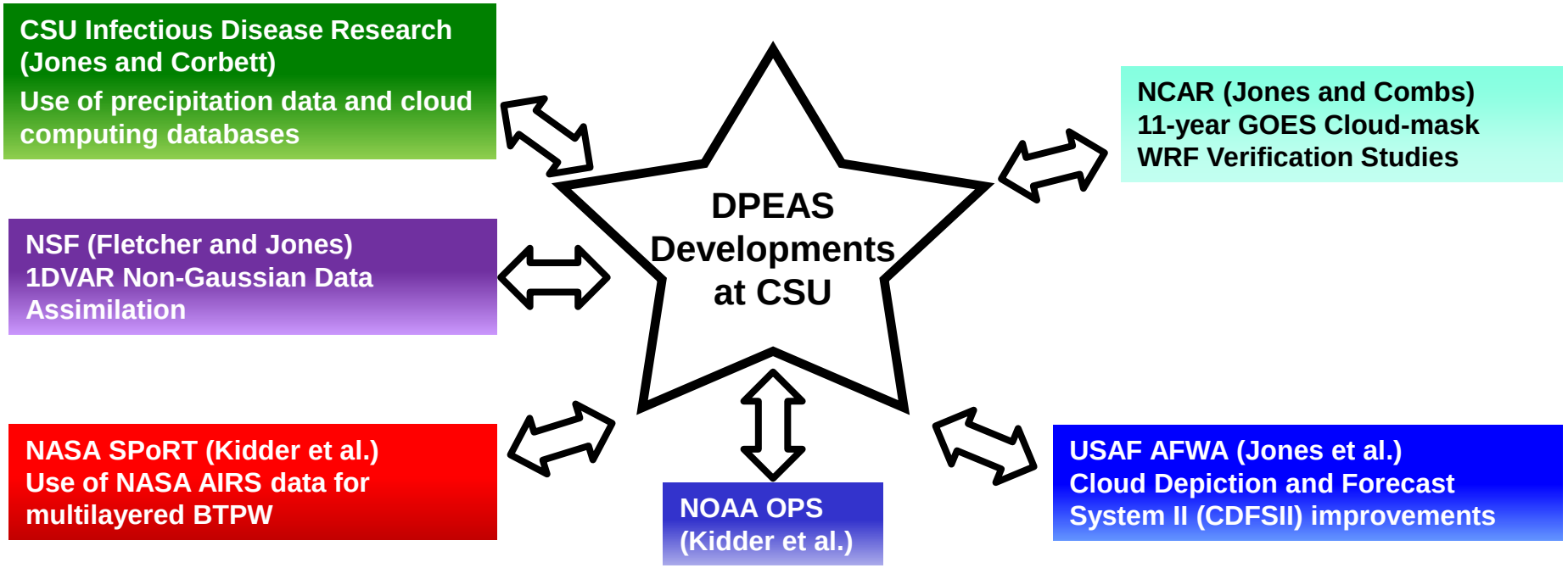
NOAA OPS (Kidder et al.)

NCAR (Jones and Combs)
11-year GOES Cloud-mask
WRF Verification Studies

USAF AFWA (Jones et al.)
Cloud Depiction and Forecast System II (CDFSII) improvements

Research to Operations: turned on its head R2O2R (+ Operations to Research)

Ongoing OPS to Research Pathway Examples using DPEAS at CSU:



- **Bottomline:** It's MUCH BETTER to have OPS dollars go further by having dual research use of the DPEAS OPS code base (and vise versa for the Research dollars). Improvements go further with each round of joint development.

Where to Next?

- The focus is on (1) enhancing the error analysis and data fusion capabilities, (2) data monitoring, and (3) on-the-fly data analysis, as DPEAS tends to be last integrative data system to touch the data before it heads out the door to forecasters
- “Corrections” and fixes are typically applied within the DPEAS data fusion algorithms
- There is a significant need for *location intelligence* (i.e., “Give me the information at this spatial location and for these time conditions”) to serve these value-added data fusion products

A New Concept in Early Demonstration Mode

1. Creation of integrated US/international sensor products
2. Emphasis on the data fusion, data assimilation, and the multisensor integration capabilities of DPEAS
3. Coupled with cloud-based delivery of location intelligent data

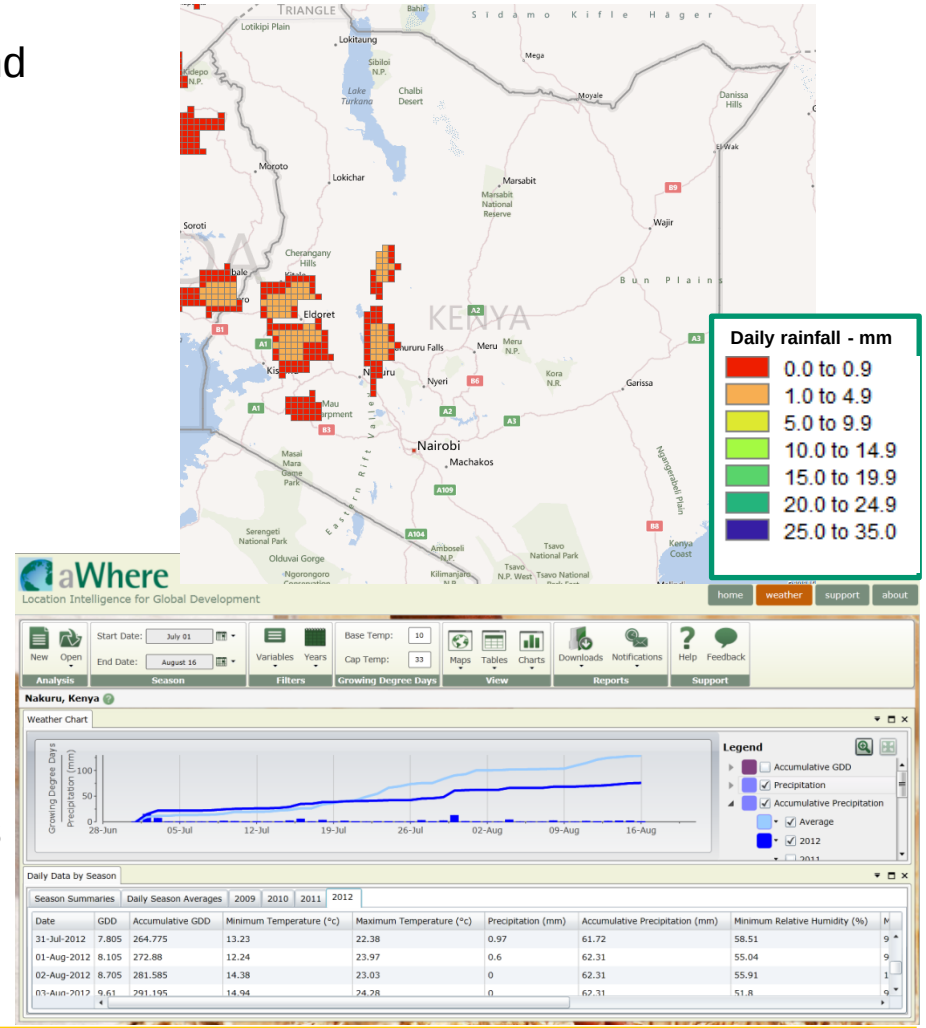
- Makes use of a scalable aWhere, Inc.[1] cloud computing database to serve data and analytics capabilities for feedback to the DPEAS system design and data analysis activities
- Designed to be interactive with large interdisciplinary science groups (e.g., Infectious Disease and Food Security)

[1]aWhere is a for-profit company with clients that range from Monsanto to Bill and Melinda Gates Foundation. aWhere serves data and a Location Intelligent SaaS platform to International Ag/Food Security/Health users requiring integrated “Big Data” and analytics capabilities

DPEAS customized for the Infectious Disease community

- A DPEAS processing algorithm and output module is customized for export to the aWhere cloud-based database
- The DPEAS output data is imported by aWhere
- The value-added data is sharable with anyone on the internet
- Can be downloaded via Excel spreadsheets or GIS tools
- The analytics feedback can be instantaneous and leverage existing satellite cal/val or other activities (stretching dollars)
- **The cool part:** The entire process can be replicated to any sensor or model data set

See more at http://me.aWhere.com



Conclusions

- DPEAS is in NOAA operational use, and the path is well-worn as an R2O OPS pathway
- R2O2R is where things are moving toward
- Systems need to stay small/modular enough to remain flexible to management needs
- New concepts for greatly expanding the reach and integration of international sensors and a diverse body of users through cloud computing will be the next wave forward
- Special thanks to our collaborators and sponsors
- Contact me at Andrew.S.Jones@ColoState.edu