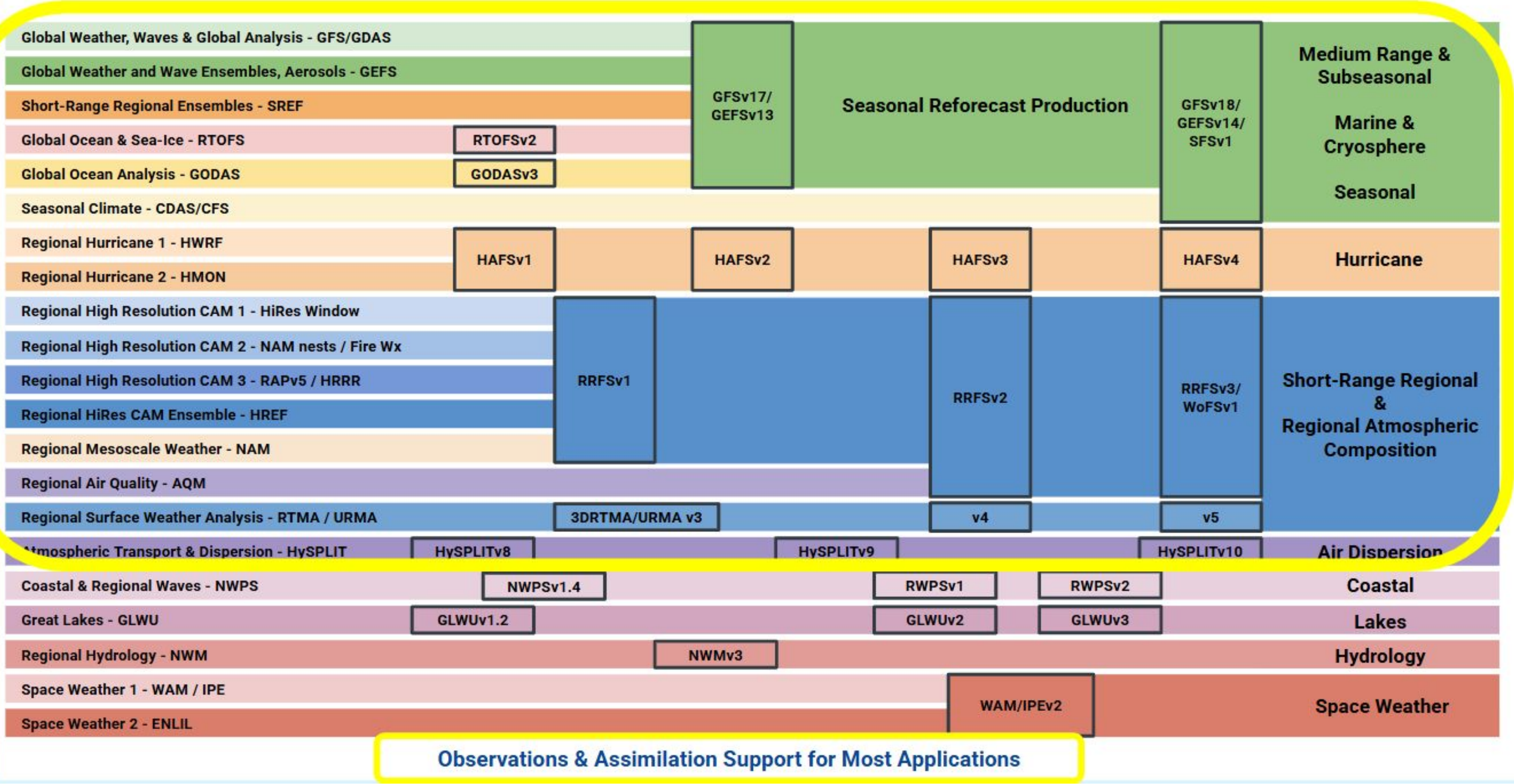
Data Assimilation Considerations from the Perspective of Operational Numerical Weather Prediction

Daryl T. Kleist & Daniel Holdaway (Presenting)
NOAA/NWS/NCEP/Environmental Modeling Center
Acting DAQC Branch Chief

14 April 2026

NATIONAL WEATHER SERVICE


Global Weather, Waves & Global Analysis - GFS/GDAS
Global Weather and Wave Ensembles, Aerosols - GEFS
Short-Range Regional Ensembles - SREF
Global Ocean & Sea-Ice - RTOFS
Global Ocean Analysis - GODAS
Seasonal Climate - CDAS/CFS
Regional Hurricane 1 - HWRF
Regional Hurricane 2 - HMON
Regional High Resolution CAM 1 - HiRes Window
Regional High Resolution CAM 2 - NAM nests / Fire Wx
Regional High Resolution CAM 3 - RAPv5 / HRRR
Regional HiRes CAM Ensemble - HREF
Regional Mesoscale Weather - NAM
Regional Air Quality - AQM
Regional Surface Weather Analysis - RTMA / URMA
Atmospheric Transport & Dispersion - HySPLIT
Coastal & Regional Waves - NWPS
Great Lakes - GLWU
Regional Hydrology - NWM
Space Weather 1 - WAM / IPE
Space Weather 2 - ENLIL
RTOFSv2
GODASv3
GFSv17/ GEFSv13
Seasonal Reforecast Production
GFSv18/ GEFSv14/ SFSv1
Medium Range & Subseasonal
Marine & Cryosphere
Seasonal
HAFSv1
HAFSv2
HAFSv3
HAFSv4
Hurricane
RRFSv1
RRFSv2
RRFSv3/ WoFSv1
Short-Range Regional & Regional Atmospheric Composition
3DRTMA/URMA v3
v4
v5
HySPLITv8
HySPLITv9
HySPLITv10
Air Dispersion
NWPSv1.4
RWPSv1
RWPSv2
Coastal
GLWUv1.2
GLWUv2
GLWUv3
Lakes
NWMv3
Hydrology
WAM/IPEv2
Space Weather
Observations & Assimilation Support for Most Applications

NWS/NCEP/EMC Review and Current State

- Introduction
- Atmospheric Systems
  - GSI Unification, Hybrid DA
- Marine, Land, Composition, and Coupled Assimilation
- Current Use of Observations
  - Decoding and Obs Processing
  - In-situ and anchor obs
  - Remotely sensed (non-satellite)
  - Remotely sensed (satellite)
- Monitoring & Observation Impacts
- Current Implementation Procedure

Current State of Data Assimilation Capabilities at NOAA/NWS/NCEP/EMC

U.S. Department of Commerce
National Oceanic and Atmospheric Administration
National Weather Service
National Centers for Environmental Prediction
5830 University Research Court
College Park, MD 20740-3818

Office Note 514
https://doi.org/10.25923/pjs0-4j42

Current State of Data Assimilation Capabilities at NCEP's Environmental Modeling Center

Daryl Kleist, Jacob R. Carley, Andrew Collard, Emily Liu, Shun Liu, Cory R. Martin, Catherine Thomas, Russ Treadon, Guillaume Vernieres

NOAA/NWS/NCEP Environmental Modeling Center

College Park, Maryland

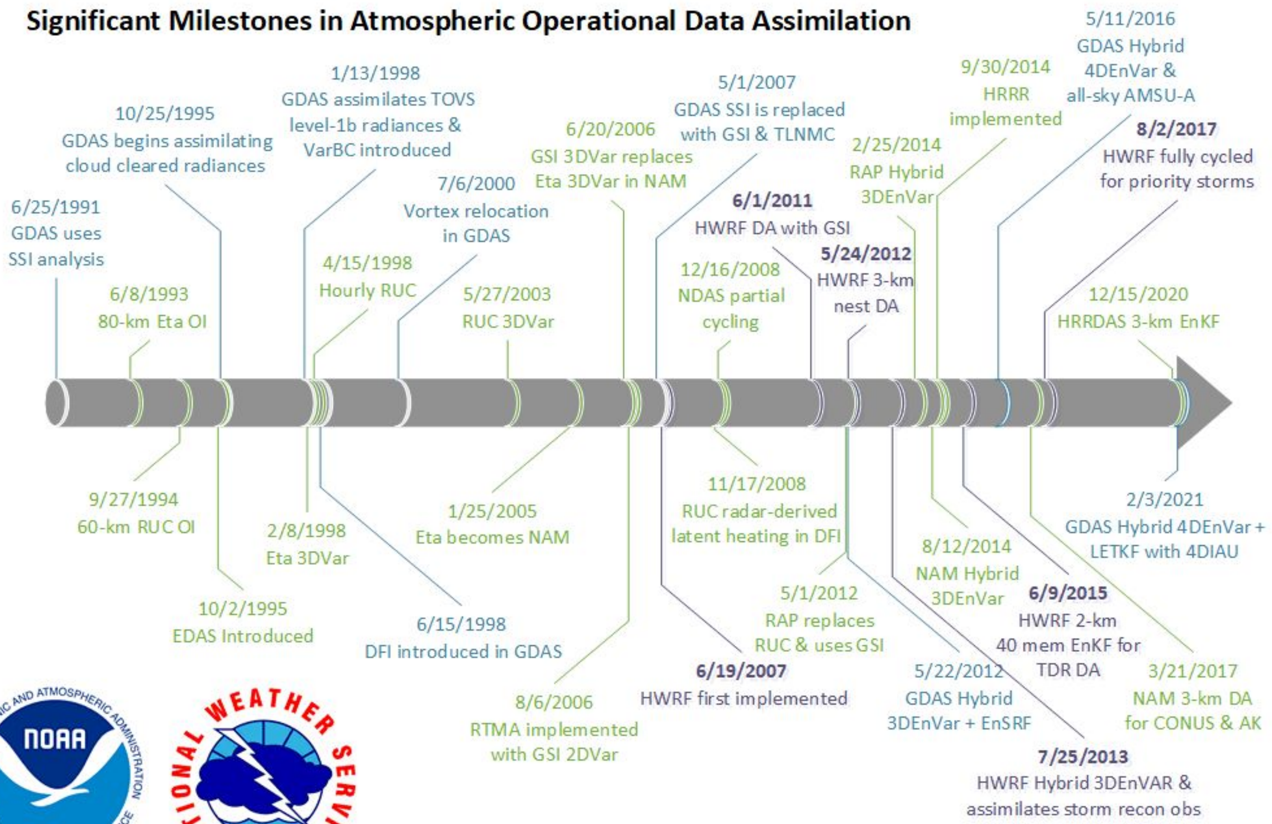
October 2023

https://doi.org/10.25923/pjs0-4j42

Historical Overview of NCEP Global DA

<table>
<tr><th>Year</th><th>1991</th><th>1995</th><th>1998</th><th>2007</th><th>2012</th><th>2014</th><th>2015</th><th>2016</th><th>2017 - 2018</th><th>2019 - 2020</th><th>2021</th><th>2022</th></tr>
<tr><th>Analysis</th><td colspan="3">Spectral Statistical Interpolation</td><td colspan="9">Gridpoint Statistical Interpolation (GSI)</td></tr>
<tr><th rowspan="2">Algorithm</th><td colspan="4">3DVar</td><td colspan="3">Hybrid, 3DEnVar
T574/T254 L64
EnSRF, 80 ensembles</td><td colspan="2">Hybrid, 4DEnVar
T574/T574 (39 km) L64
EnSRF, 80 ensembles</td><td>Hybrid, 4DEnVar
T766/T766 (25 km) L64
EnSRF, 80 ensembles</td><td colspan="2">Hybrid, 4DEnVar
T766/T766 (25 km) L127
LETKF, 80 ensembles</td></tr>
<tr><td colspan="8"></td><td colspan="4">Near Sea Surface Temperature (NSST) Analysis - 3DVar for foundation temperature since 2017</td></tr>
<tr><th>Balance</th><td colspan="3"></td><td colspan="9">Tangent-linear Normal Model Constraint (TLNMC)</td></tr>
<tr><th>Satellite Data</th><td>NESDIS Retrievals</td><td colspan="7">● Direct Radiance Assimilation
● Clear-sky Radiance Assimilation</td><td colspan="4">● Direct Radiance Assimilation
● Correlated obs error for Hyperspectral IR sensors (2021)
● All-sky Framework for MW sensors
○ All-sky Assimilation for AMSU-A (2016) and ATMS (2019)
○ Non-precipitating cloud affected radiances are assimilated
○ Cloudy scenes are assumed to be overcast</td><td>● Direct Radiance Assimilation
● Correlated obs error for Hyperspectral IR
● All-sky Framework for ATMS/AMSU-A
○ Precipitation-affected data assimilated
○ Fractional cloud coverage accounted for
○ Consistent use of cloud microphysics assumptions between model and DA</td></tr>
<tr><th>Obs Operator</th><td></td><td colspan="2">Community Radiative Transfer Model (CRTM)</td><td colspan="4">CRTM Scattering Solver
Advanced Doubling & Adding Method (ADA)</td><td colspan="5">CRTM with Scattering Solver (ADA) + Simulation under Fractional Cloud Coverage (2017)
● Cloud Overlap (Total Cloud Cover)
● Two-column Radiance Calculation</td></tr>
<tr><th>Bias Correction</th><td colspan="2">Offline</td><td colspan="3">Two-step Bias correction
● VarBC for air-mass predictors
● Time-moving average of O-F for scan angle predictors</td><td colspan="7">Variational Bias Correction (VarBC) for all predictors</td></tr>
<tr><th rowspan="2">Forecast</th><td colspan="2">Spectral Model
Sigma Eulerian
T126
L18 (1991)
L28 (1993)</td><td colspan="5">Spectral Model
Hybrid Semi-Lagrangian (2007)
Zhao-Carr Microphysics(cloud water and ice)
T170 (105km) L42 (1998)
T574 (23km) L64 (2010)</td><td colspan="2">Spectral Model
Hybrid Semi-Lagrangian
Zhao-Carr Microphysics
T1534 (13km)
L64, Model Top 55km</td><td>FV3 Dynamic Core
GFDL Microphysics
5 Hydrometers
T1534 (13km)
L64, Model Top 55km</td><td colspan="2">FV3 Dynamic Core
GFDL Microphysics
5 Hydrometers
T1534 (13km)
L127, Model Top 80km</td></tr>
<tr><td colspan="6"></td><td colspan="6">Stochastic Physics Parameterization for Ensembles since 2015: SPPT (tendency), SHUM (moisture), and SKEB (winds)</td></tr>
<tr><th>Initialization</th><td colspan="3"></td><td colspan="6">Digital Filter Initialization (DFI)</td><td>No DFI</td><td colspan="2">4D Incremental Analysis Update (IAU)</td></tr>
</table>

History of Atmospheric DA



What are some of the biggest drivers (and challenges) for operational DA/NWP?

Operational Constraints

- **Operational NWP can be broken down into four primary steps:**
  1. Observation collection, decoding, and pre-processing
  2. Data Assimilation
  3. Model Integration
  4. Post-processing & product dissemination
- **Schedule:** Very strict on-time delivery requirements, fixed (pre-determined, reliable schedule)
  1. NWS/NCEP/EMC: ~25 minutes for operational DA
  2. ECMWF: ~45 minutes
- This necessitates compromise between science/theory and computational issues and practicalities

Operational GFS/GDAS
6 hourly cycling, -/+ 3 hour observation window

00 UTC GFS

EVENT	Average Start Time	Average End Time	STATUS	COMMENTS
DATA DUMP AND PREP	02:37:08	02:59:49	COMPLETE-03:00:16	ON-TIME
ANALYSIS	03:00:01	03:26:50	COMPLETE-03:27:12	ON-TIME
T1534 FORECAST F000-F384	03:27:02	05:14:04	COMPLETE-05:13:44	ON-TIME
12hr PRODUCTS	03:41:39	03:45:12	COMPLETE-03:44:43	ON-TIME
24hr PRODUCTS	03:46:24	03:48:27	COMPLETE-03:49:56	ON-TIME
36hr PRODUCTS	03:49:39	03:51:34	COMPLETE-03:53:05	ON-TIME
48hr PRODUCTS	03:52:44	03:54:44	COMPLETE-03:55:58	ON-TIME
60-72hr PRODUCTS	03:55:52	04:00:55	COMPLETE-04:01:22	ON-TIME
84-120hr PRODUCTS	04:02:04	04:13:36	COMPLETE-04:15:02	ON-TIME
000-384hr GFS WAVE PRODUCTS	03:33:35	05:12:54	COMPLETE-05:12:33	ON-TIME
GFS MOS FORECAST	04:13:36	04:13:49	COMPLETE-04:15:22	ON-TIME

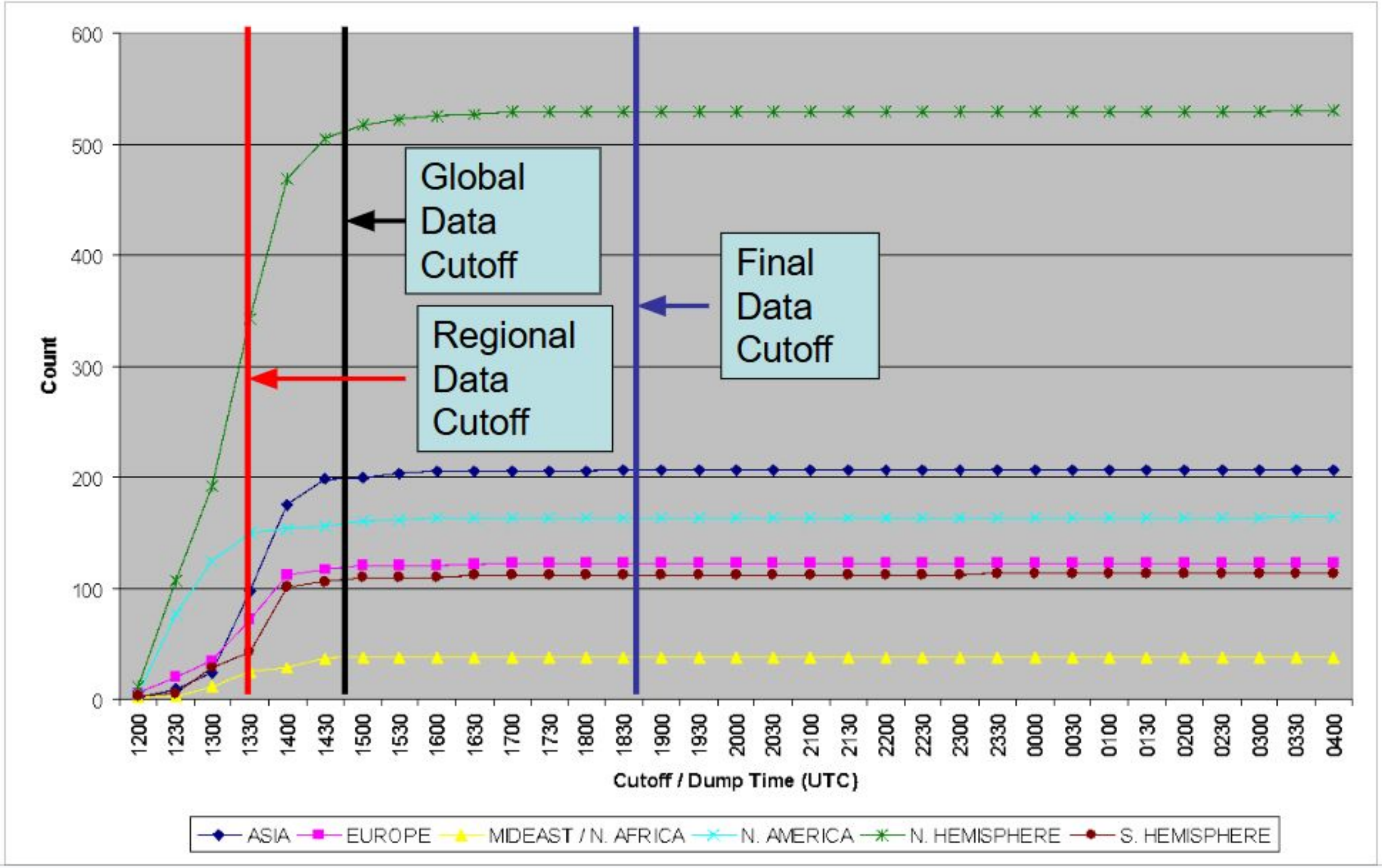
00 UTC GDAS

EVENT	Average Start Time	Average End Time	STATUS	COMMENTS
DATA DUMP AND PREP	05:40:08	06:02:26	COMPLETE-06:03:25	ON-TIME
ANALYSIS	06:02:39	06:40:11	COMPLETE-06:41:31	ON-TIME
FORECAST	06:47:50	07:08:25	COMPLETE-07:09:52	ON-TIME

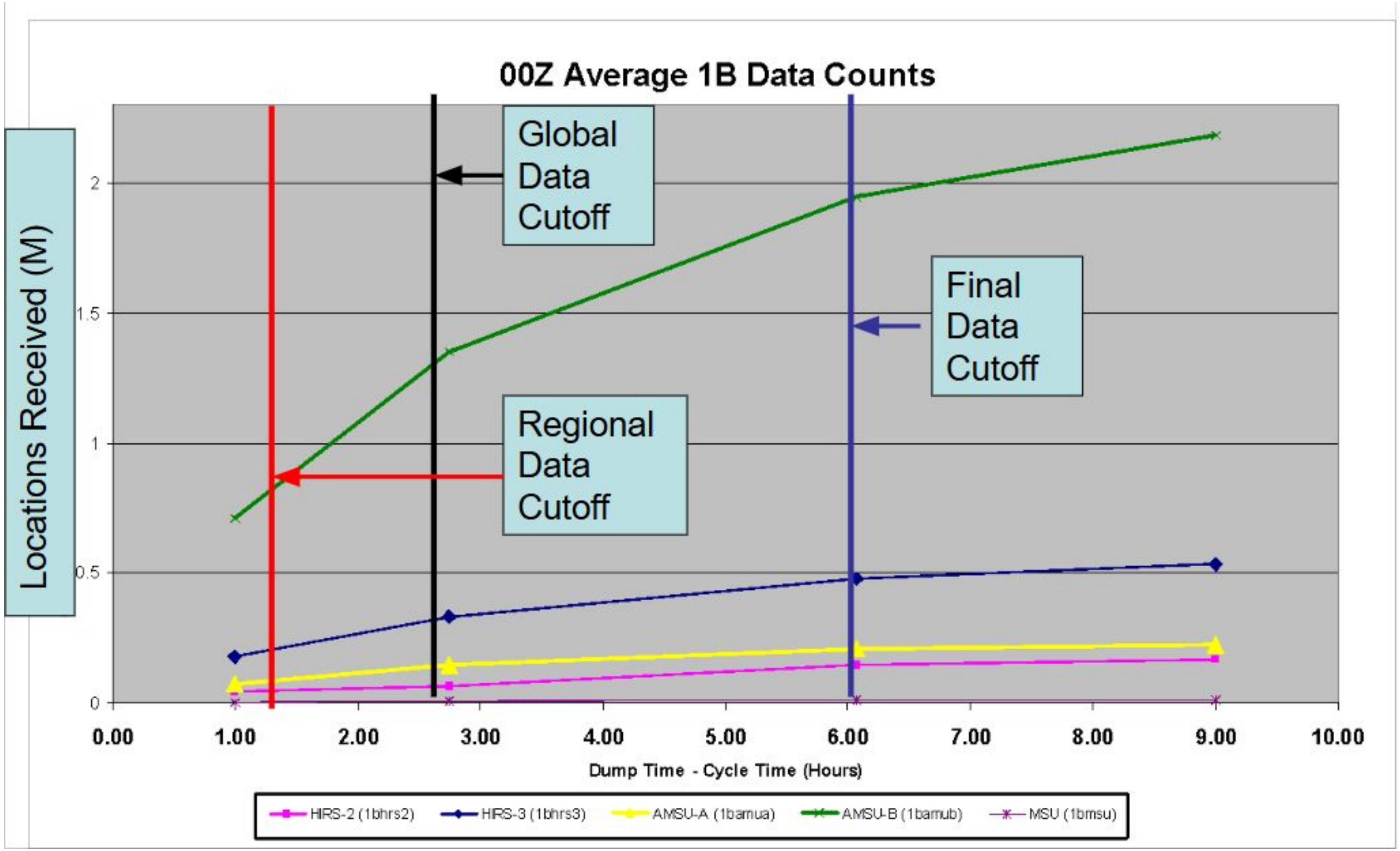
Why two cycles for the same analysis time for a global model?

Any data not available by cut-off will not be used
Later “catch-up” cycle at +5:40

Rawinsondes



Legacy POES Data Delivery



Challenge: data latency

Issue for real-time assimilation: many observation arrive 1-3 hours after their valid time

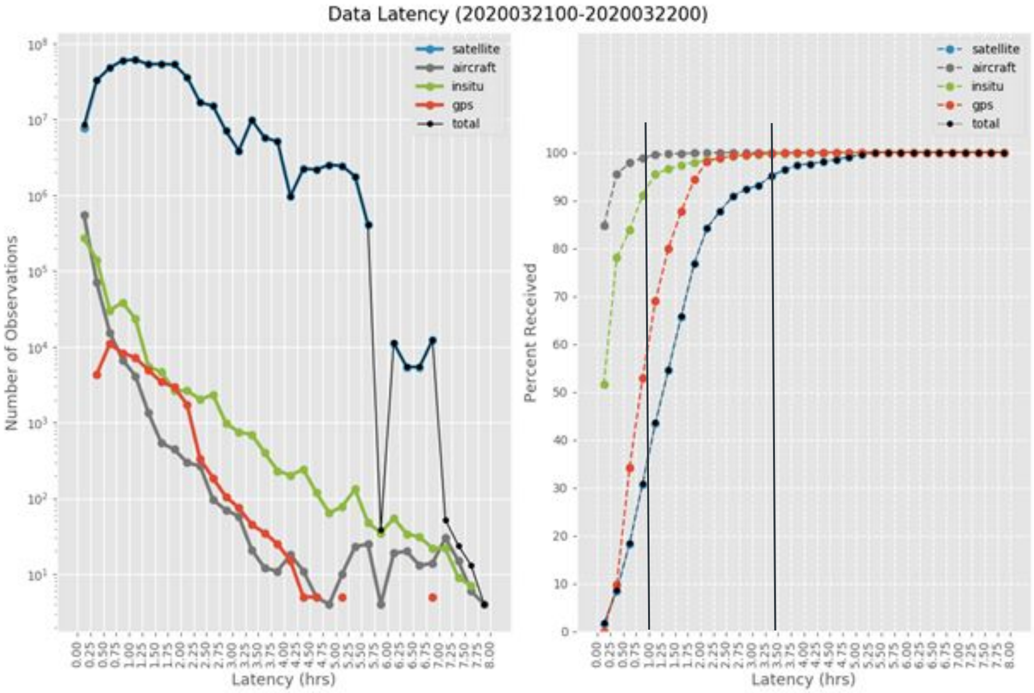
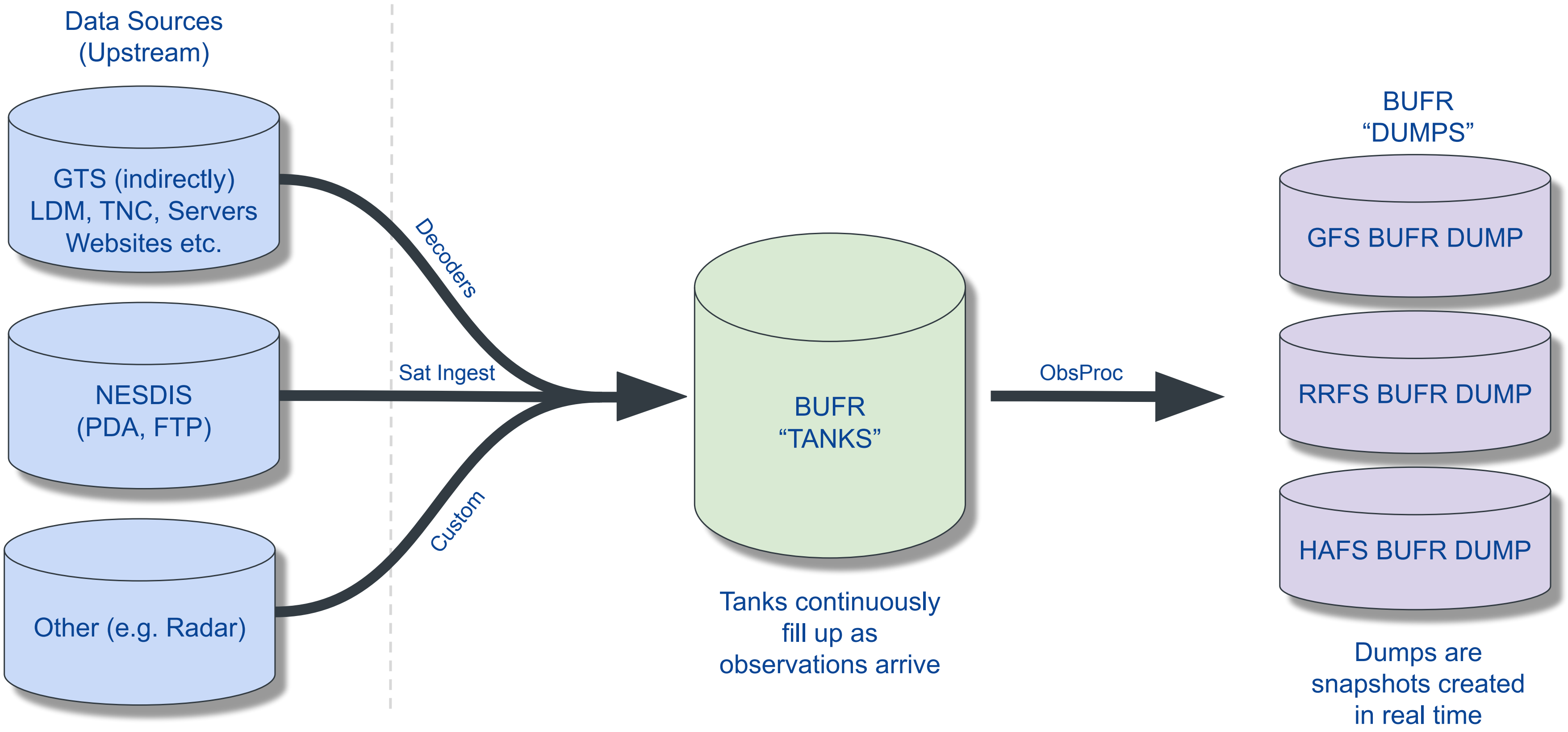


Figure courtesy D. Lippi

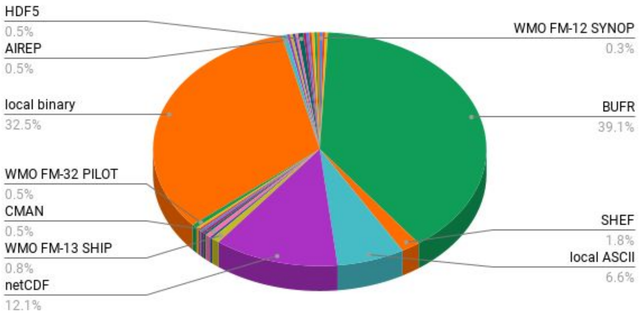
Data latency = receipt time - valid time

Example:
An observation from a radiosonde that is valid at 00Z might not be available until 01Z. By this time, it would already have missed the cutoff time in a system with a [-0.5hr, 0.5hr] data window

Introduction: Where Does Observations Processing (ObsProc) Fit?

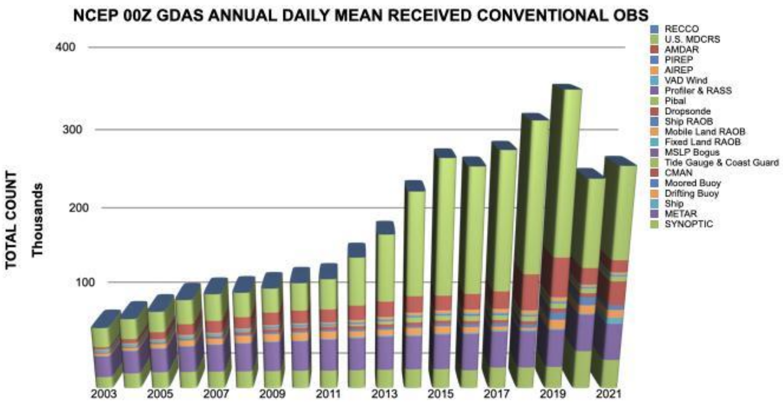


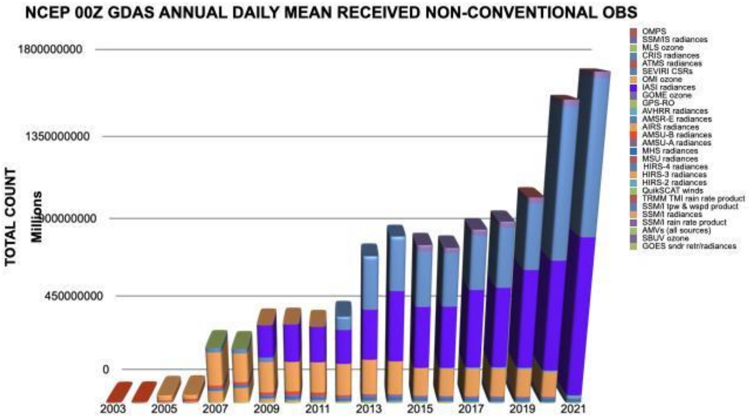
Volumes of received and processed observations



⇐ Receiving 385 data types, 29 data formats

- ~500 dump files per day for GDAS and GFS
- ~4880 files for day for all regional models and CDAS





Larger data volumes expose “technical debt” and present engineering challenges

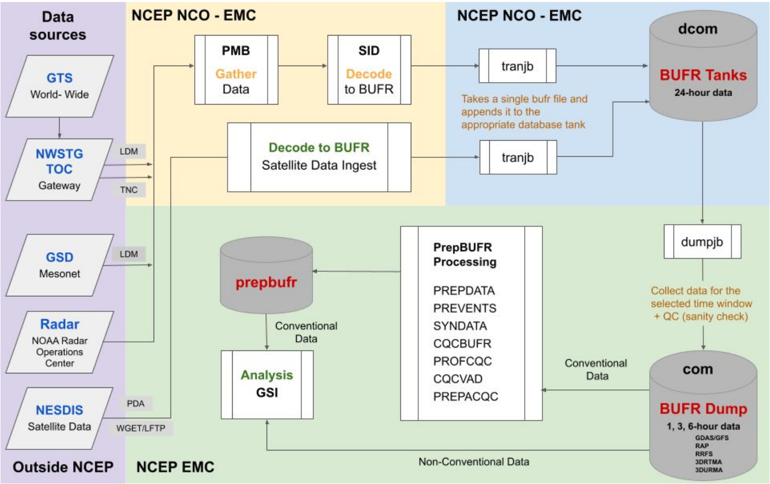
ObsProc Re-engineering

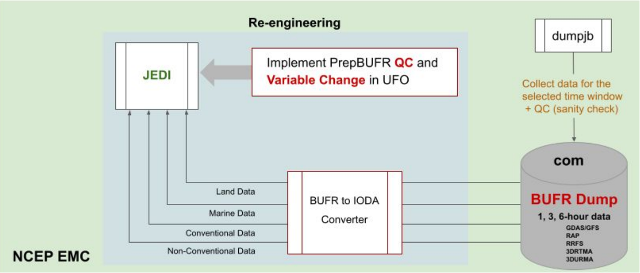
- NCEP Analysis (GSI) will be transitioned into JEDI
- **ObsProc re-engineering is underway to reform and adopt the changes**:
    - Adopt - JEDI IODA Data Design: HDF5 and data layout
        - Develop BUFR to IODA converter
        - Follow IODA conventions: unified variable names, units, and dimensions
    - Reform - Implement PrepBUFR quality control and variable change procedures in JEDI UFO

BUFR Converter for JEDI without any changes in the decoding, tanking and dumping processes

- C++ based tool with Python API
- Supports various BUFR formats: PrepBUFR, NCEP and WMO BUFR
- Brings ObsProc from Fortran 66 to modern software development paradigms

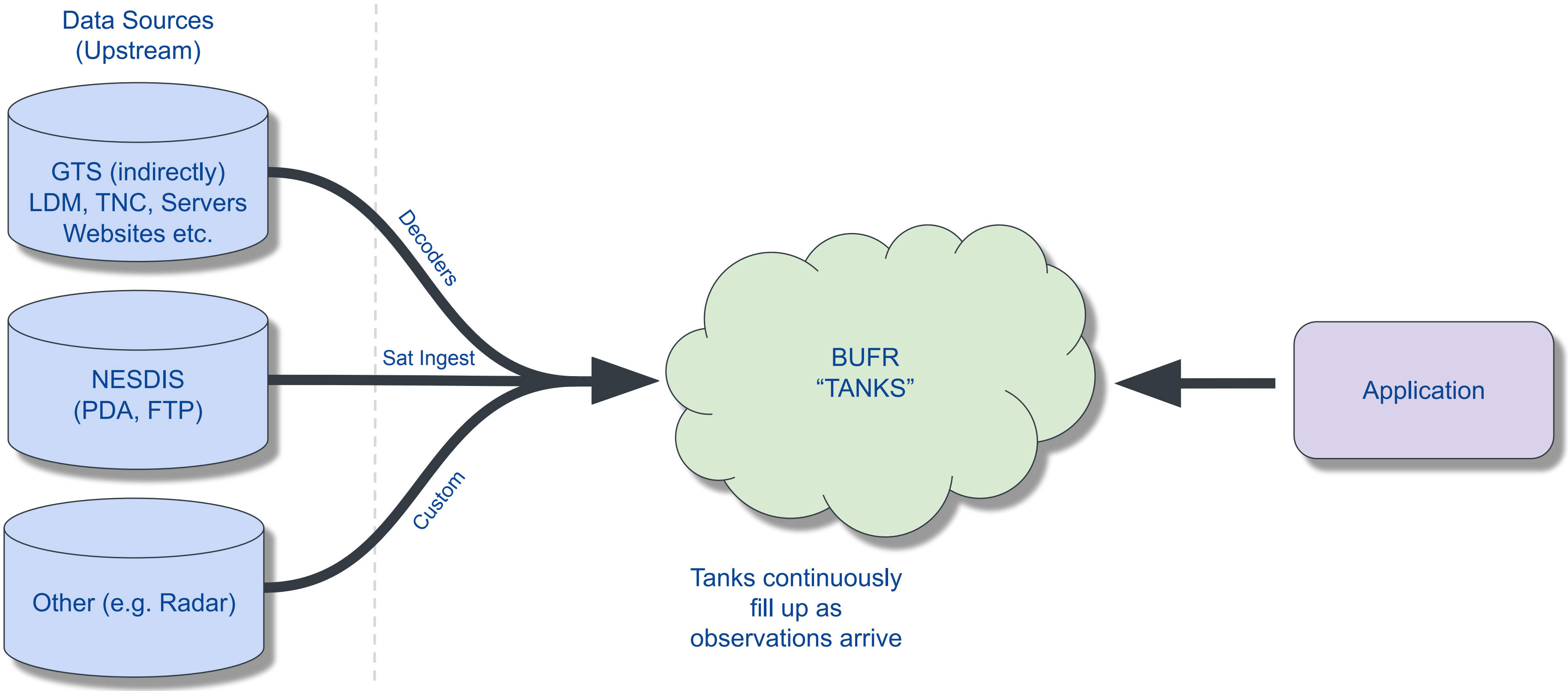
ObsProc





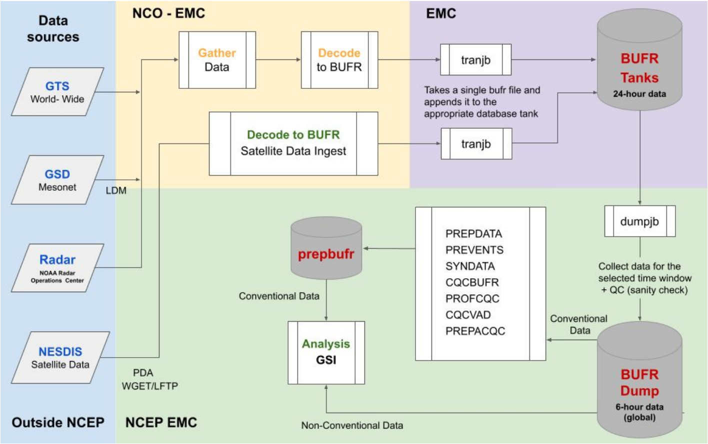
Re-engineered part of ObsProc

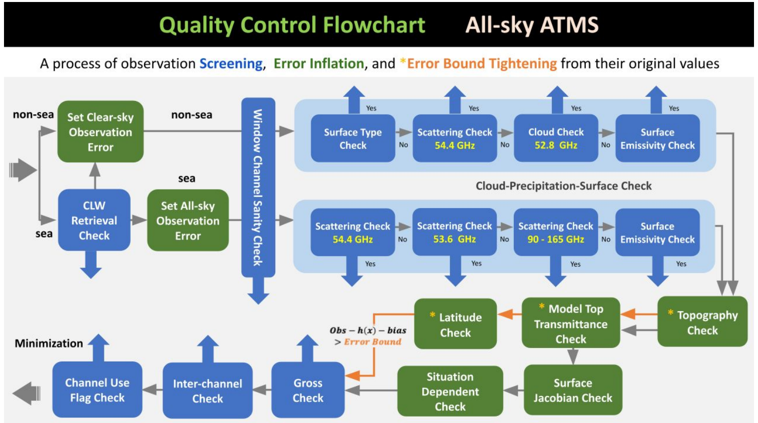
ObsProc Re-engineering - Long Term



What it takes to use an observation in NWP

1. Acquisition/ingest – including data formats, latency/timeliness, etc.
2. Observation Operator – can we map the model state to observation space (everything from interpolation to radiative transfer)? This may include TL/AD versions of operators for some solvers (e.g. variational).
3. Characterization – errors (sensor and representativeness), data quality, etc.
4. Develop application-relevant quality control, error specification, and other algorithmic changes (perhaps bias correction, etc.)
5. Testing and optimization within context of cycled application

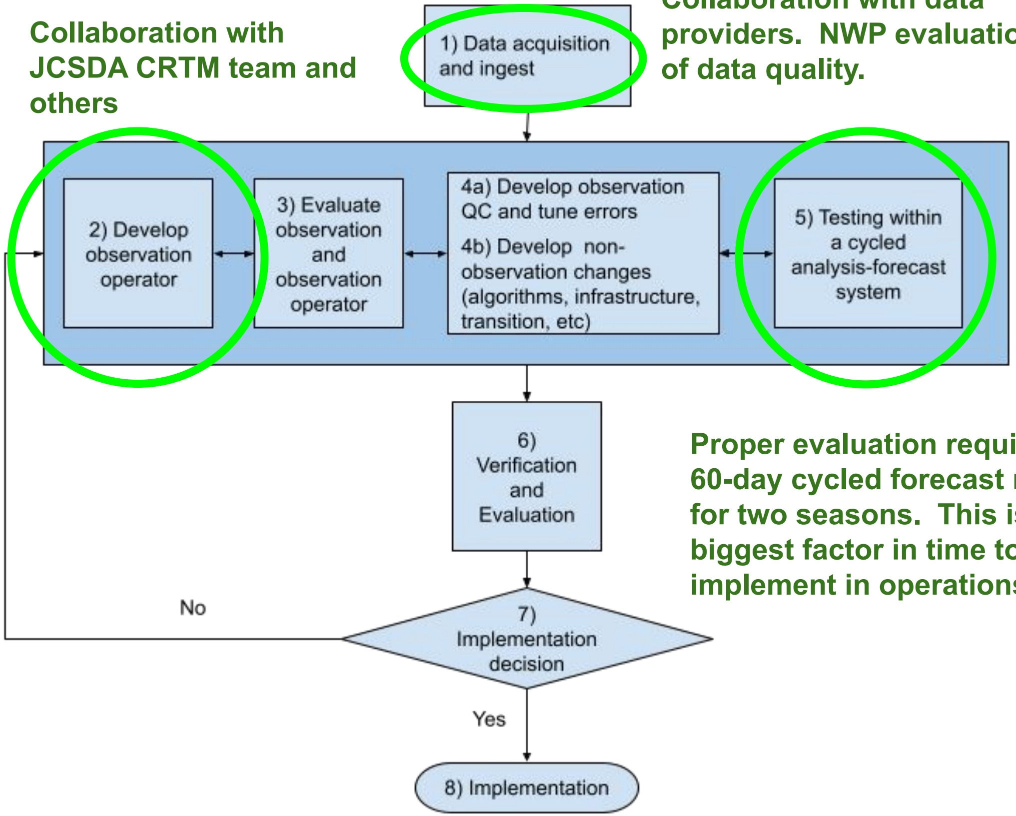




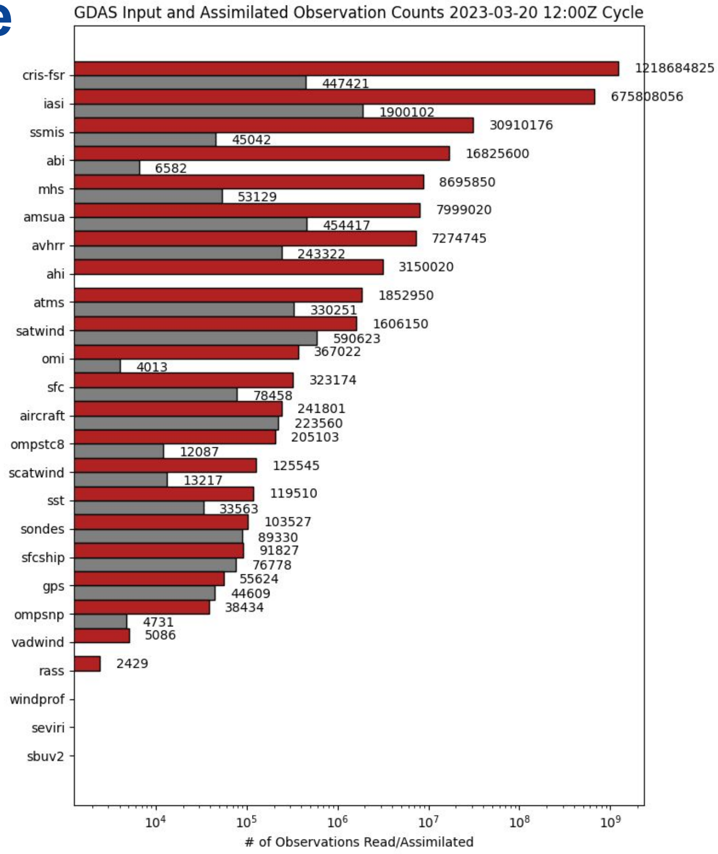
Observation Implementation Procedure

Collaboration with JCSDA CRTM team and others

Collaboration with data providers. NWP evaluation of data quality.



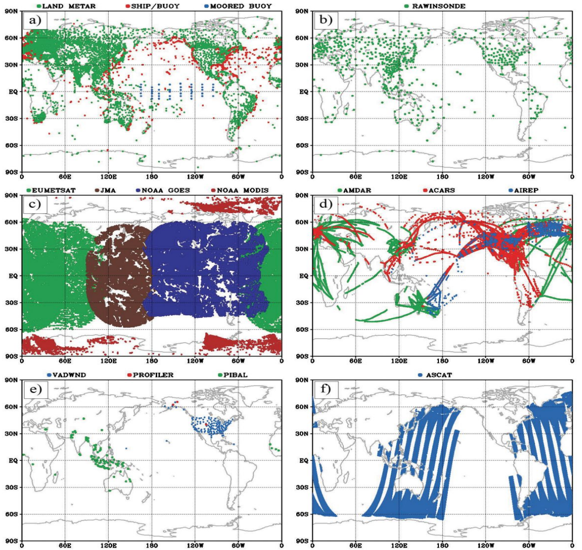
Proper evaluation requires 60-day cycled forecast runs for two seasons. This is the biggest factor in time to implement in operations.

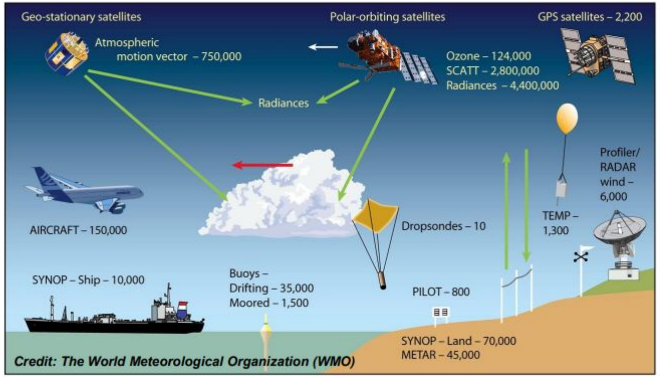


Total number of observations available to (red) and used by (gray) the GSI for each observation type for the GDAS analysis cycle valid March 20, 2023 12 UTC. Note the x-axis is on a logarithmic scale

Global Observing System

- Heterogeneous mix of measurements at widely varying spatial and temporal resolutions
- Quantities that generally are indirect measurements of predicted/analyzed variables





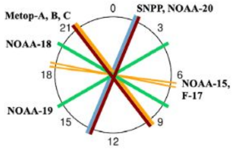


Figure 7. Equatorial crossing time of various polar-orbiting satellites that are currently operational. The orbits continue to drift, and the equator times shown are valid for mid-2018 (Source: Bormann et al., 2022, ECMWF).

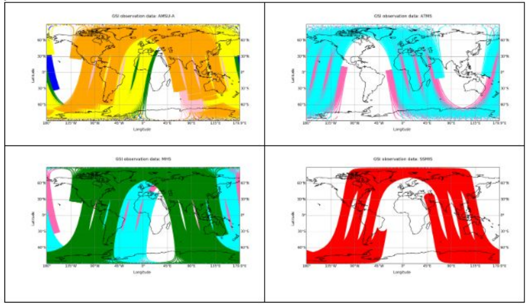
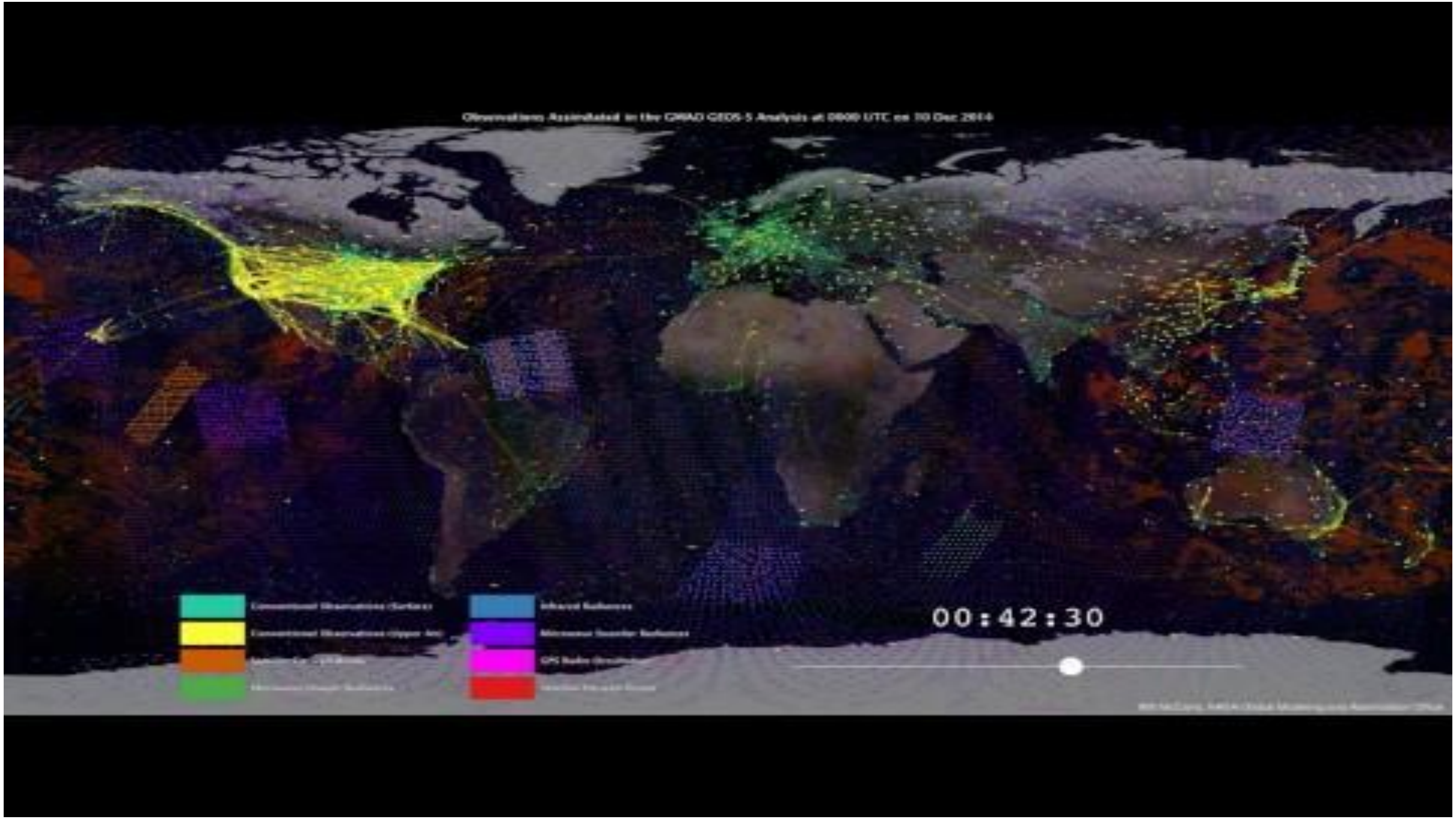


Figure 8. Spatial coverage for MW data in a 6-hour window.

Single Cycle



Satellite DA / Forward Modeling

- **One of the biggest developments in operational DA for NWP was to allow for observations to be different from analysis variables**
  1. Move away from retrievals, direct assimilation
  2. Allows for use of observations as observed *and* use of analysis variables with better properties

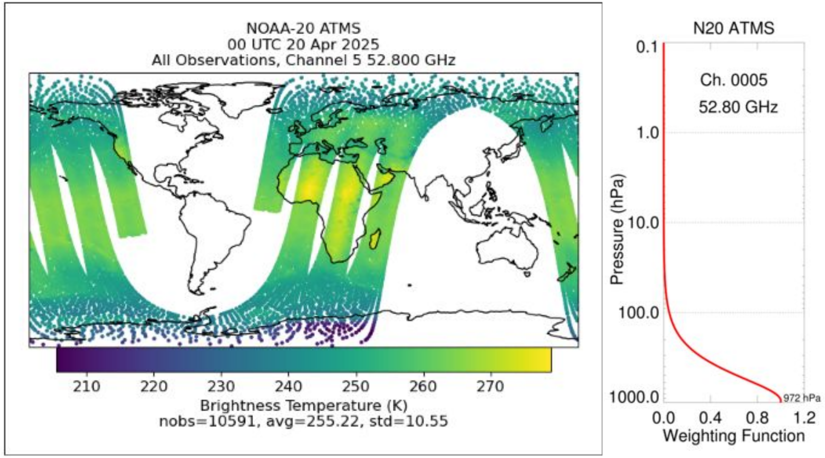
- **Radiances/Brightness Temperatures:**
  1. Indirect observations of meteorological parameters
  2. Compute radiance equivalent from temperature, humidity, surface pressure, wind, ozone (and/or other tracers)
     i. Radiative Transfer Model (and Surface Emissivity Model)

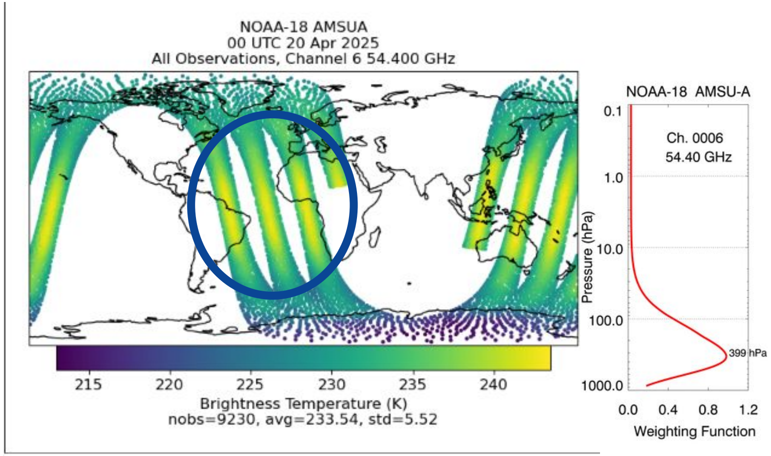
- **Some advantages**
  1. observation error specification, quality control, less introduction of auxiliary information, improved monitoring

Satellite DA / Radiative Transfer

- **In addition to simulated radiance (or brightness temperature), produce Jacobians for assimilation**

$$\frac{\partial R}{\partial T}, \frac{\partial R}{\partial q}, \frac{\partial R}{\partial q}, \frac{\partial R}{\partial O_3}, \ldots$$

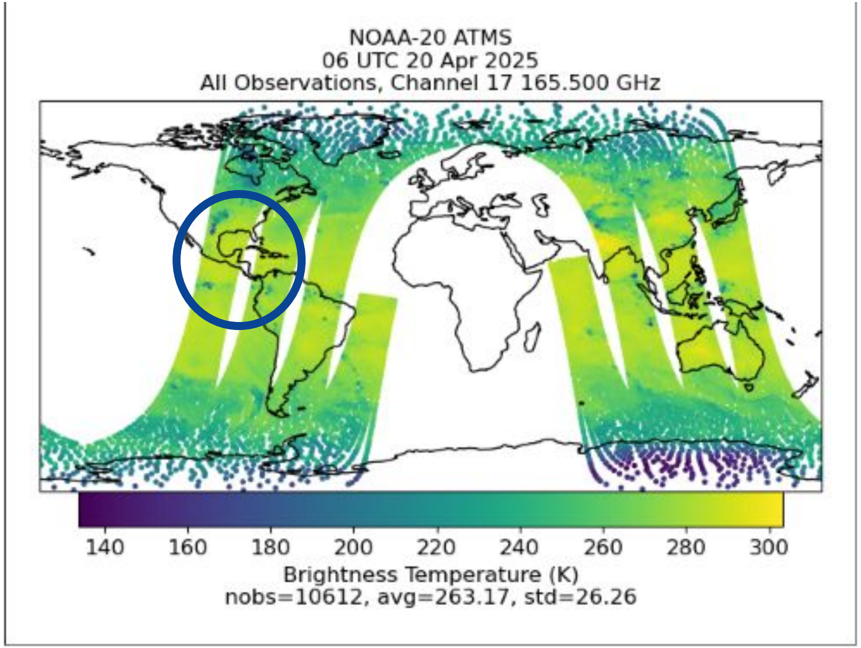


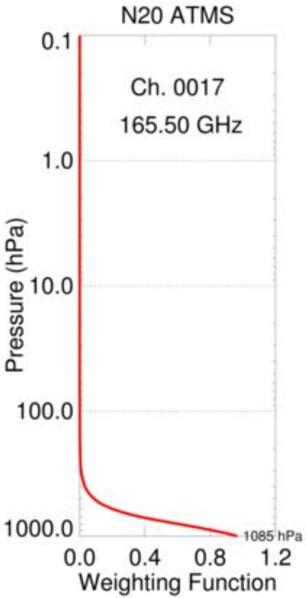


- **What to do about this feature?**

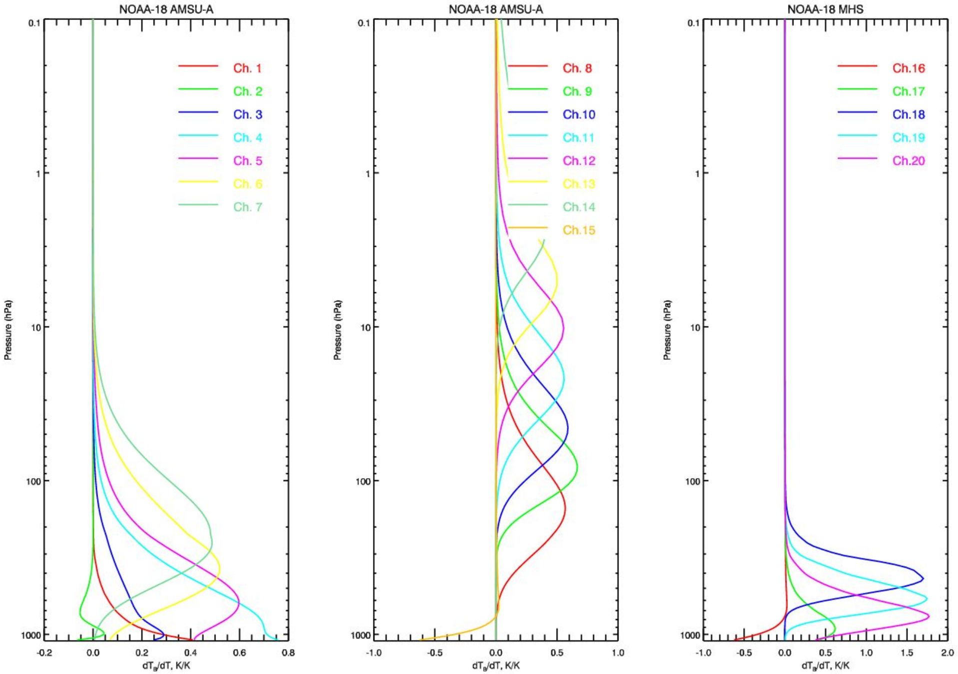
Satellite DA / Radiative Transfer

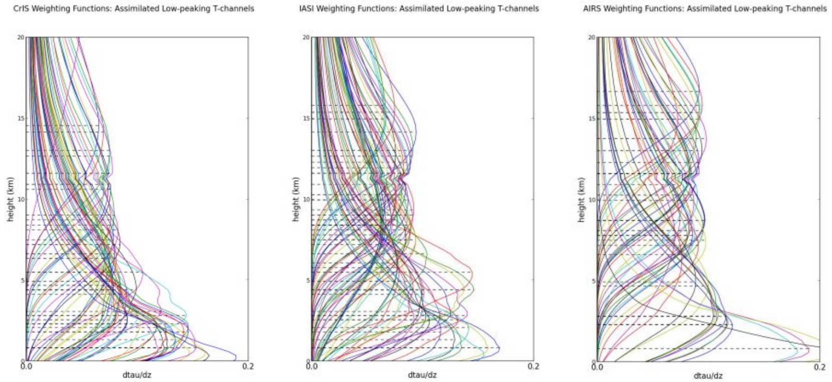
- **What might this channel be sensitive to?**





Example Weighting Functions





Courtesy UKMO

Broad in the vertical; not point measurements!

Quality Control

- **Critical for useful assimilation**
- **Data must be removed that has gross errors or that cannot be properly simulated by forward model**
- **Three general categories**
  1. Instrument problems
  2. Simulation errors (especially from things like clouds, precipitation, etc.)
  3. Surface emissivity simulation errors
- **General procedures (significant progress on all of the below toward all-sky/all-surface assimilation)**
  1. IR cannot really penetrate clouds, cloud height can be difficult to determine
  2. MW impacted by clouds/precipitation, but signal can be modeled and accounted for
  3. Surface emissivity and temperature characteristics not well known for land/ice/snow

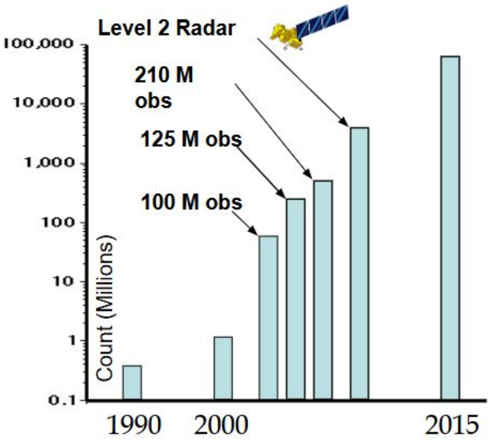
Thinning (and Super-obbing)

- **Some data is thinned and/or super-obbed prior to assimilation**
    1. Includes both horizontal thinning and channel selection (e.g. for hyperspectral IR)

- **Three primary reasons**
    1. Significant redundancy in data (radiances, AMVs)
    2. Reduced correlated observation error (AMVs)
    3. Computational Expense (Radiances)

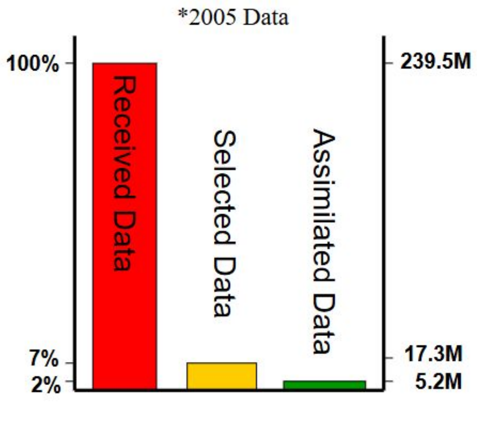
QC & Thinning: Net Result

Daily Satellite & Radar Observation Count

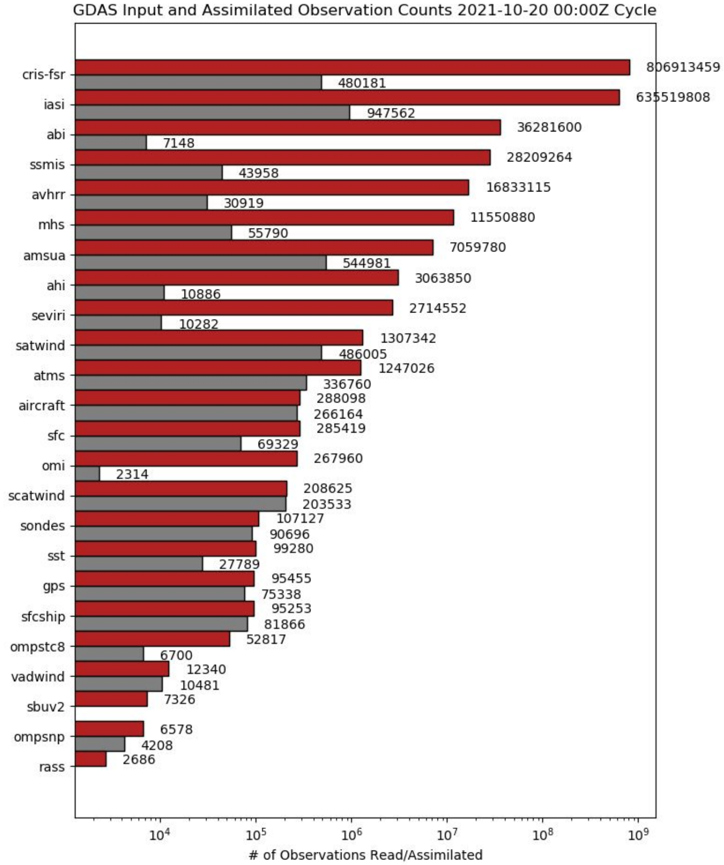


Five Order of Magnitude Increases in Satellite Data Over Fifteen Years (2000-2015)

Daily Percentage of Data Ingested into Models

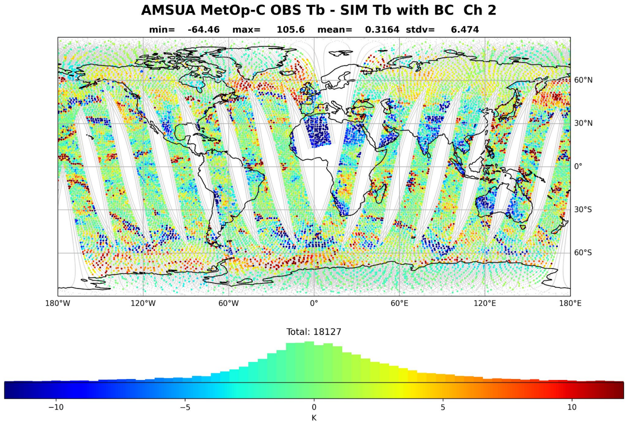


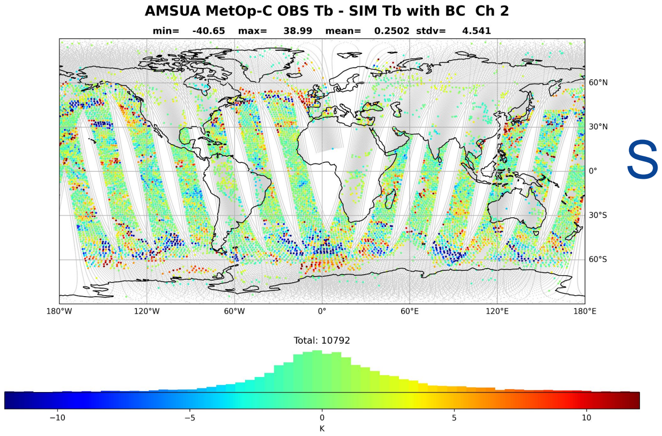
Received = All observations received operationally from providers
Selected = Observations selected as suitable for use
Assimilated = Observations actually used by models



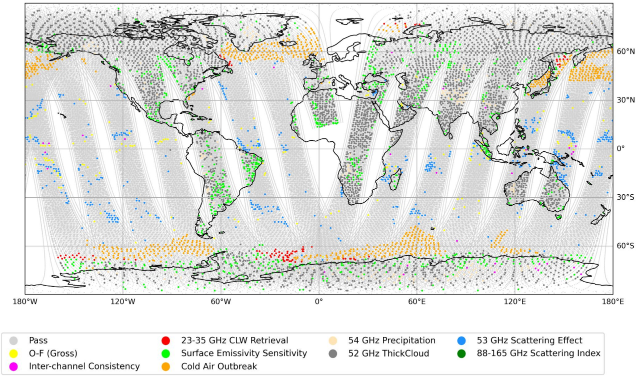
QC & Thinning: Net Result

Available





Selected



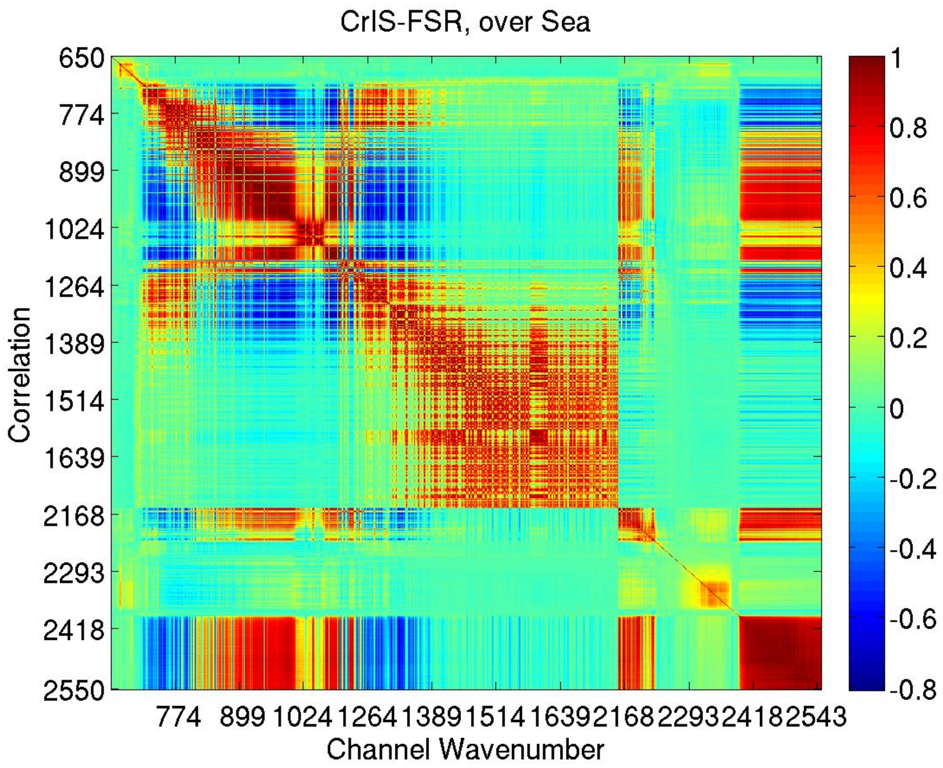
QC

Observation Errors

- **Specified based on instrument errors and innovation statistics**
- **Historically, has been used to account for other deficiencies in system**
  1. Correlated observations errors not prescribed or not well known
- **Bias must be accounted for; can be bigger than signal!**
  1. Bias Observation
  2. Inadequacies in characterization of instrument
  3. Forward model deficiencies
  4. Bias in background (!)

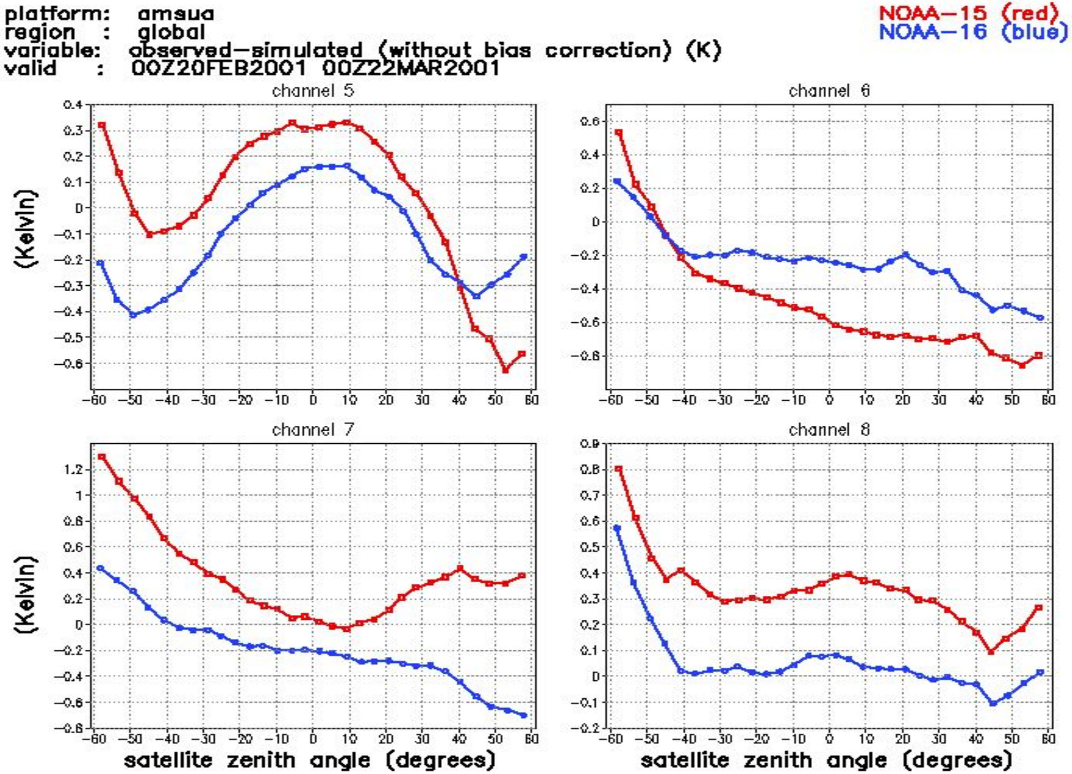
Observation Errors

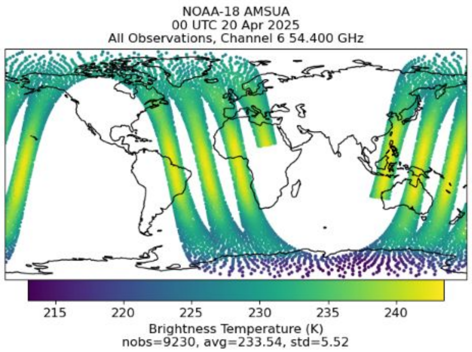
- **Work progressing toward estimation and use of correlated observation errors**



- **Why might spatially correlated observation errors be difficult?**

Bias Example





Bias Correction

- Variational Bias Correction has been critical for advanced/improved use of satellite observations (extended to other observation types)
  - Extremely effective - deals with instrument drift, representativeness issues, etc.
  - However, makes observations "look" like background in absence of anchor observations

Linear predictor model for bias in each channel:

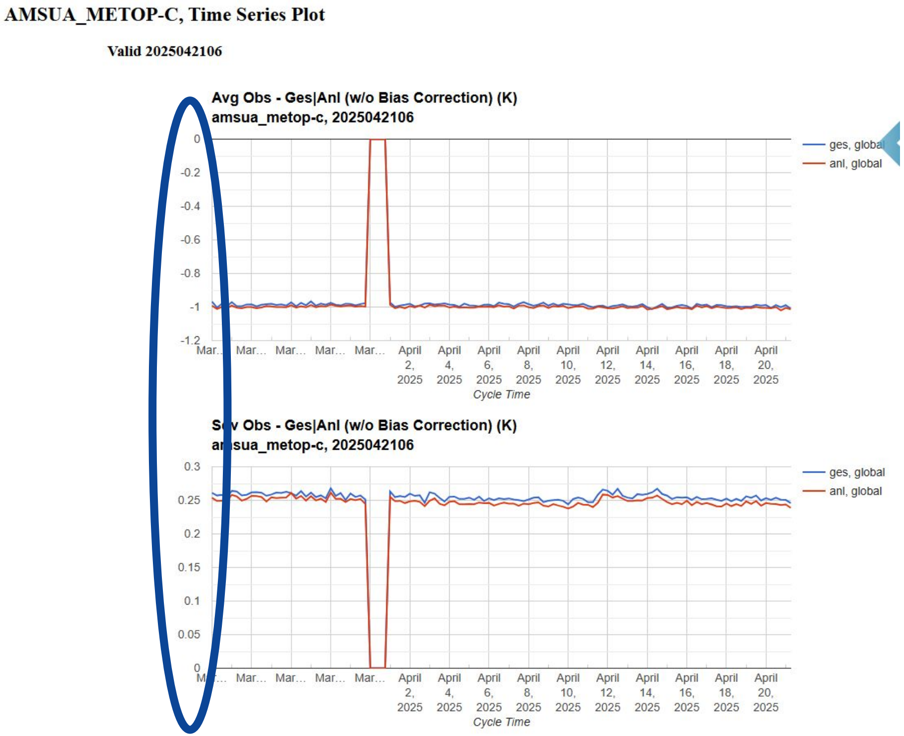
$$\mathbf{b}(\mathbf{x}, \beta) = \sum_{i=0}^{N_p} \beta_i \mathbf{p}_i(\mathbf{x})$$

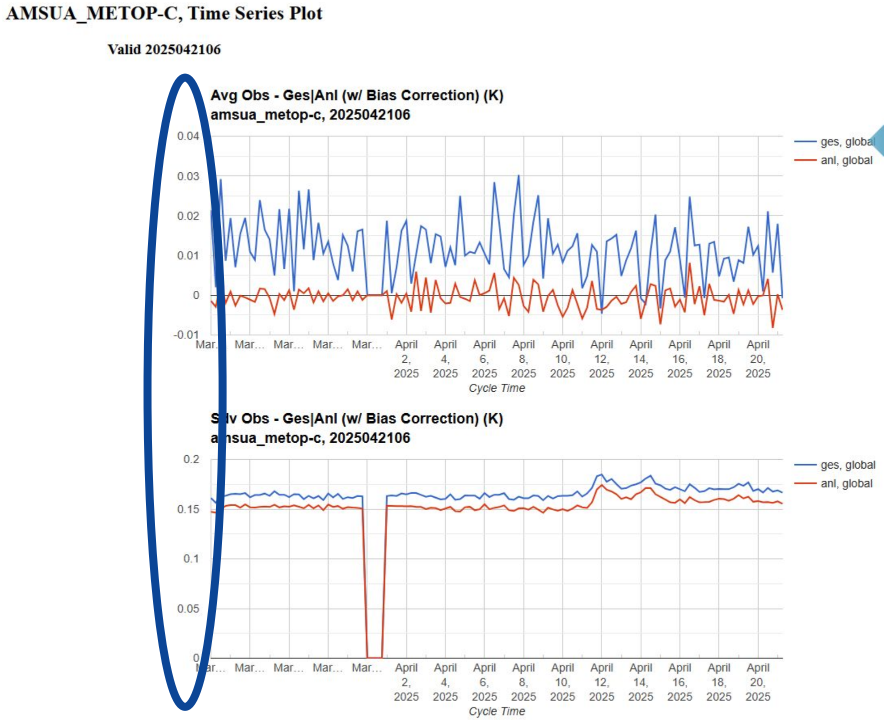
Cost function:

$$J(\mathbf{x}, \boldsymbol{\beta}) = \overbrace{(\mathbf{x}_b - \mathbf{x})^T \mathbf{B}_x^{-1} (\mathbf{x}_b - \mathbf{x})}^{J_b\text{: background constraint for x}} + \overbrace{(\boldsymbol{\beta}_b - \boldsymbol{\beta})^T \mathbf{B}_\beta^{-1} (\boldsymbol{\beta}_b - \boldsymbol{\beta})}^{J_\beta\text{: background constraint for } \beta}$$

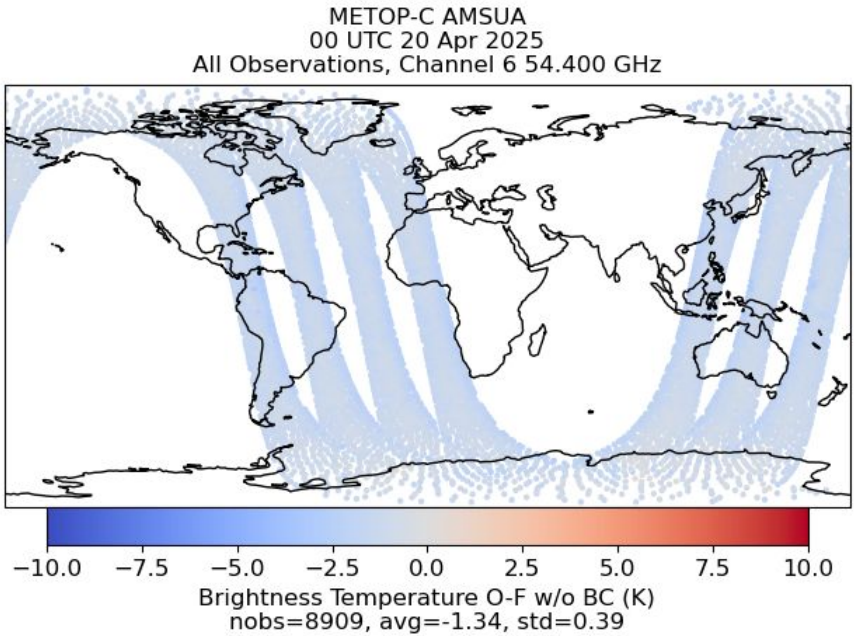
$$+ \underbrace{[\mathbf{y} - \mathbf{b}(\mathbf{x}, \beta) - h(\mathbf{x})]^T \mathbf{R}^{-1} [\mathbf{y} - \mathbf{b}(\mathbf{x}, \beta) - h(\mathbf{x})]}_{J_o\text{: bias-corrected observation constraint}}$$

Bias Correction Example

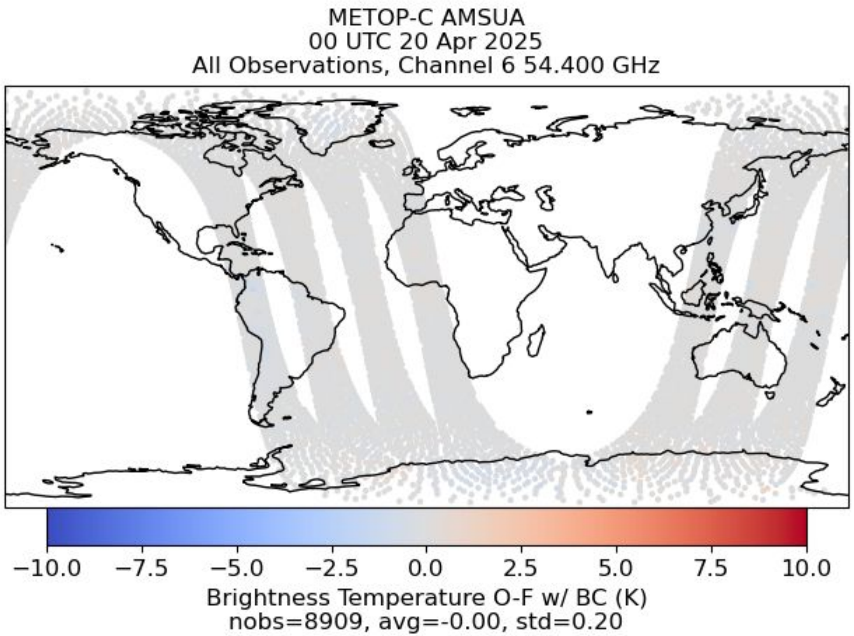




Bias Correction Example

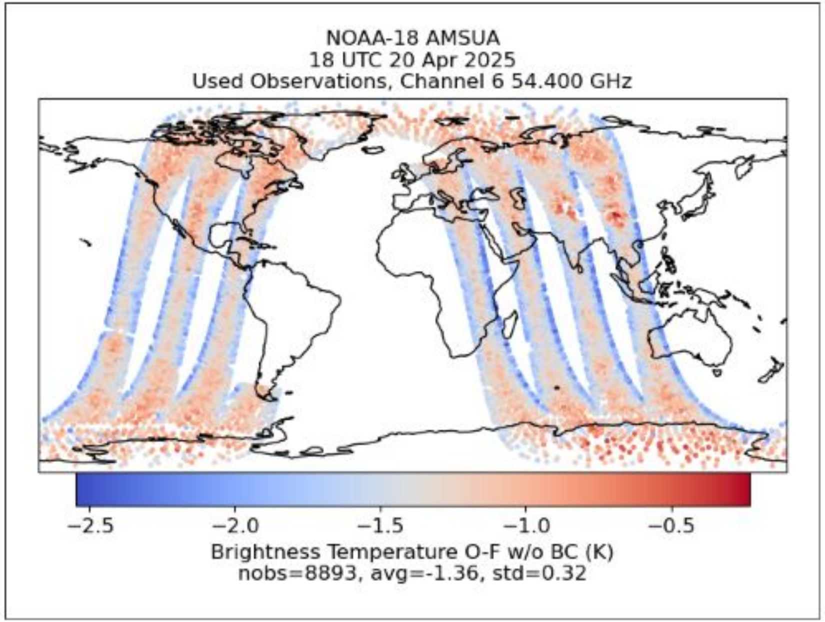


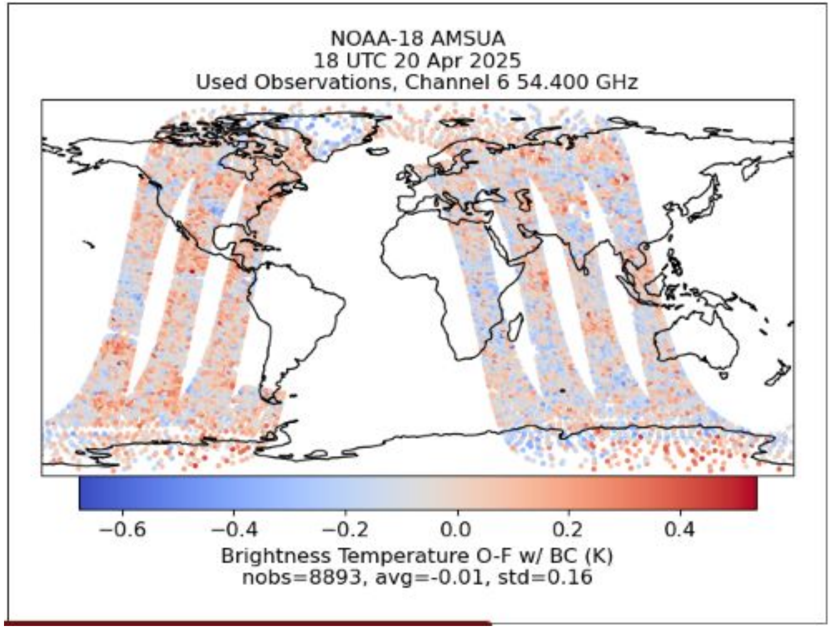
O-F, NoBC



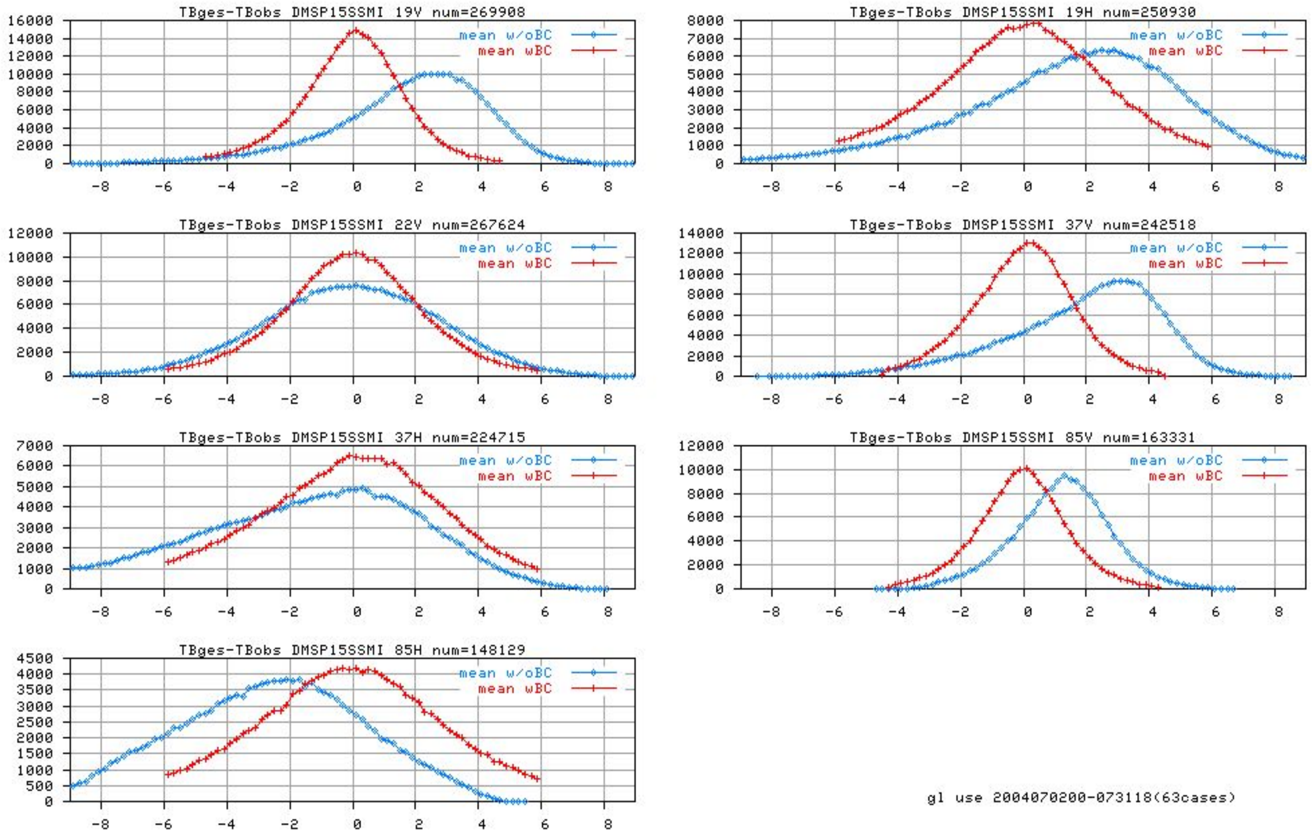
O-F, BC

Bias Correction Example



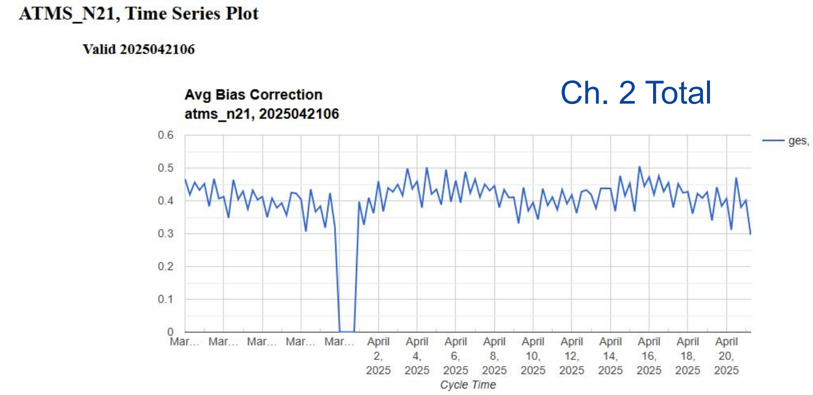


Bias Correction Example

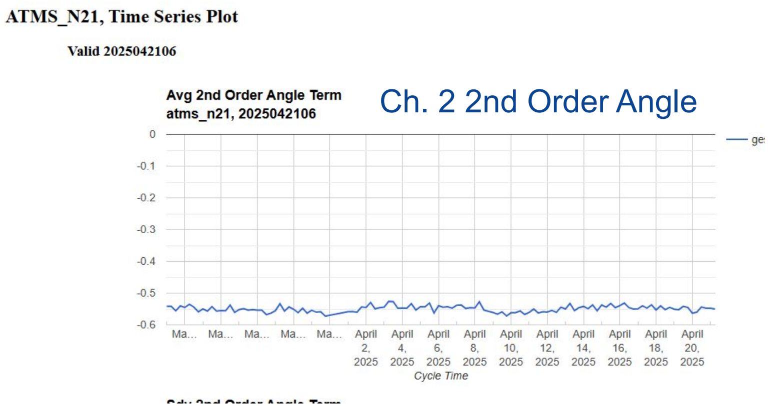


Bias Correct: Evolves in Time

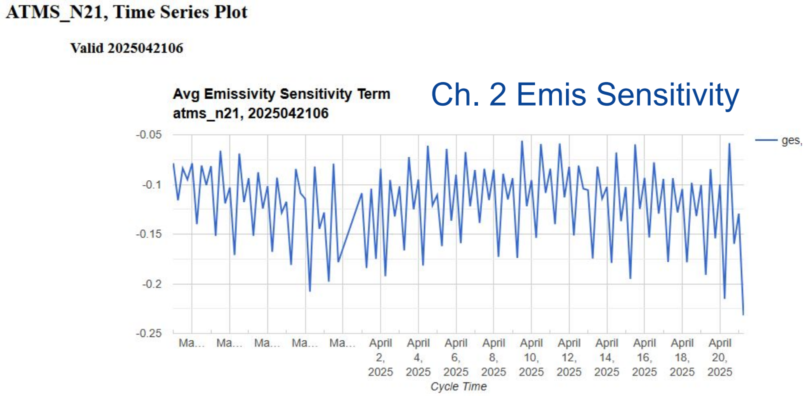
Ch. 2 Total



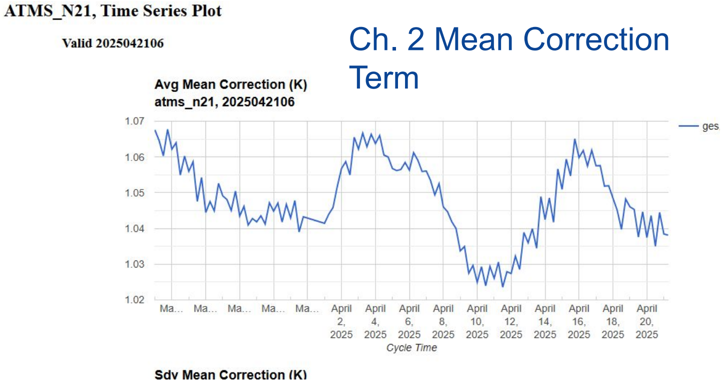
Ch. 2 2nd Order Angle



Ch. 2 Emis Sensitivity

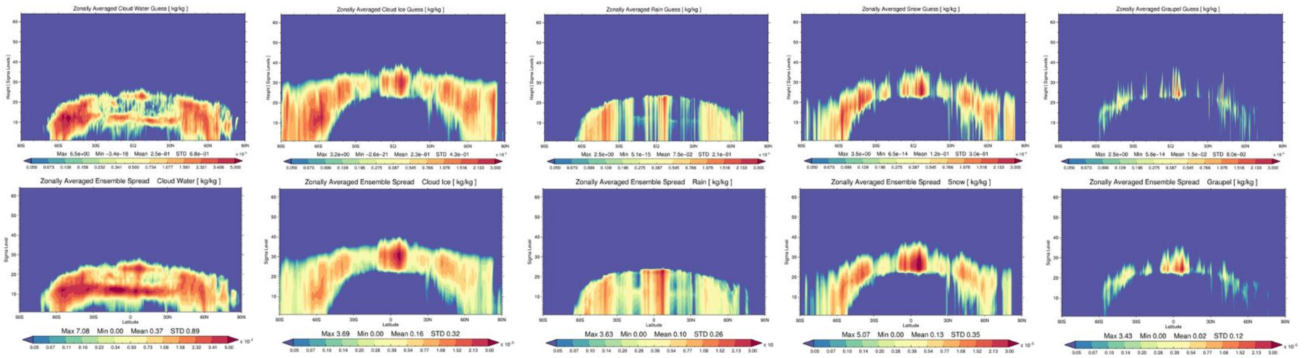


Ch. 2 Mean Correction Term



What Makes the MW Cloudy Radiance Assimilation Possible?

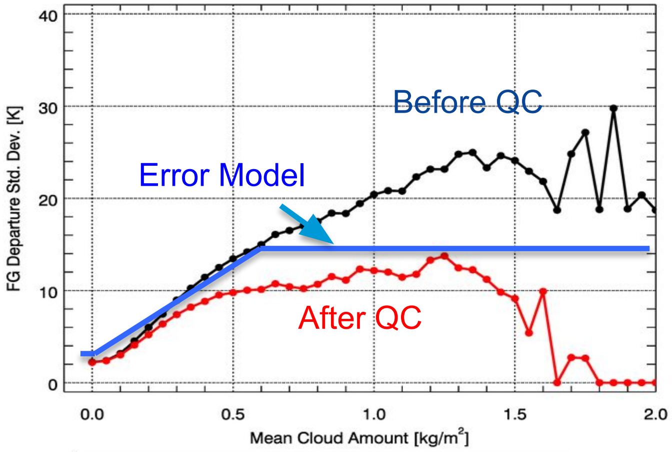
- **Improvement in forecast model**
  - Resolution - higher horizontal and vertical, raised model top
  - Representation of cloud and moisture processes
  - Initialization technique
- **Advanced data assimilation techniques**
  - Flow-dependent background error covariance through the use of ensemble
  - Representation of model uncertainties through stochastic physics
  - Situation dependent observation error model
- **More accurate radiative transfer modeling**
  - Scattering by hydrometers
  - Sub-grid variability for cloud and precipitation overlap
  - Simulation under fractional cloud coverage

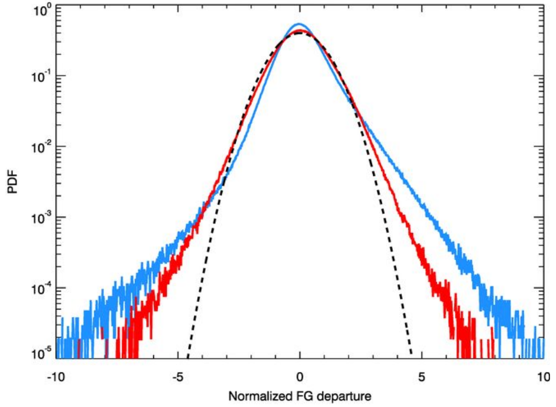


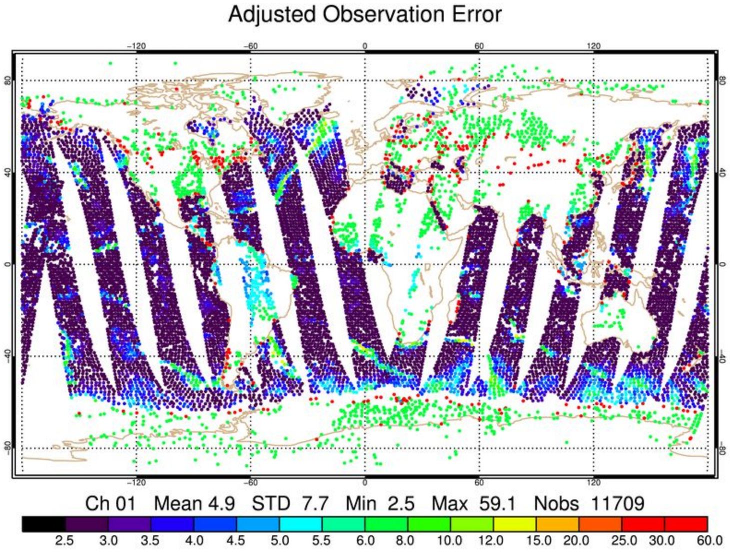
Zonal Averaged Hydrometer Mixing Ratio

Zonal Averaged Hydrometer Ensemble Spread

All-sky Framework – Symmetric Observation Error

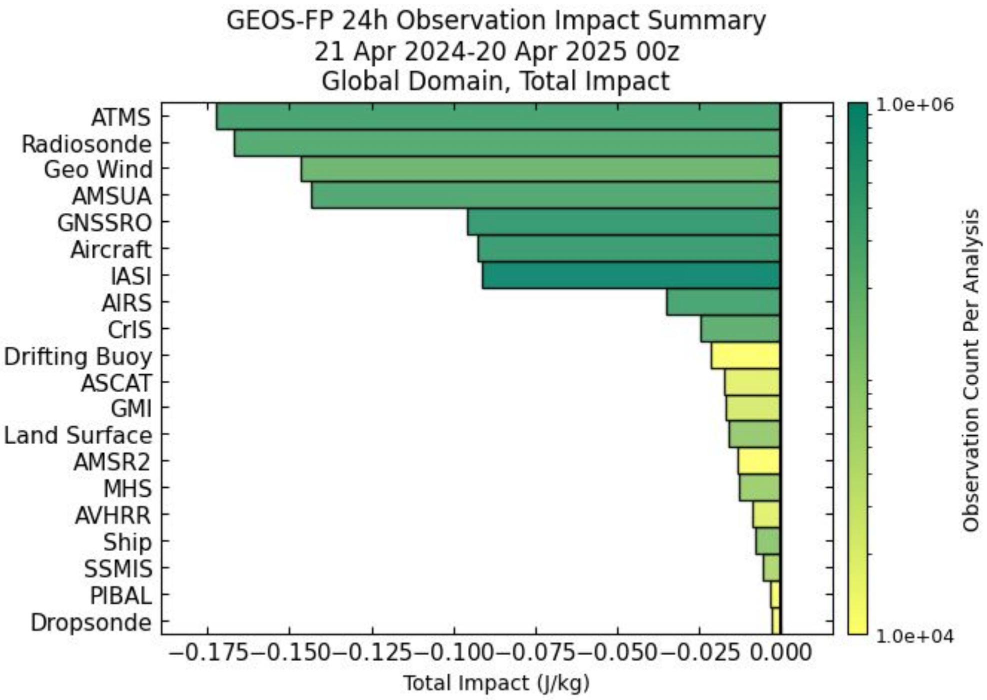


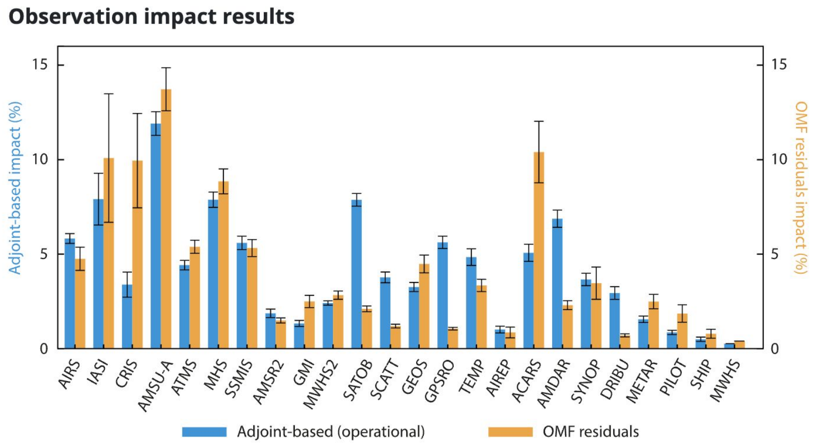




- Observation error is assigned as a function of the symmetric cloud amount (Geer et al. 2011)
- Gross check ±3 of the normalized FG departure (accept Gaussian part of obs)

Why so much time covering radiances?





Courtesy
ECMWF

https://gmao.gsfc.nasa.gov/forecasts/systems/fp/obs_impact/

Operational Constraints

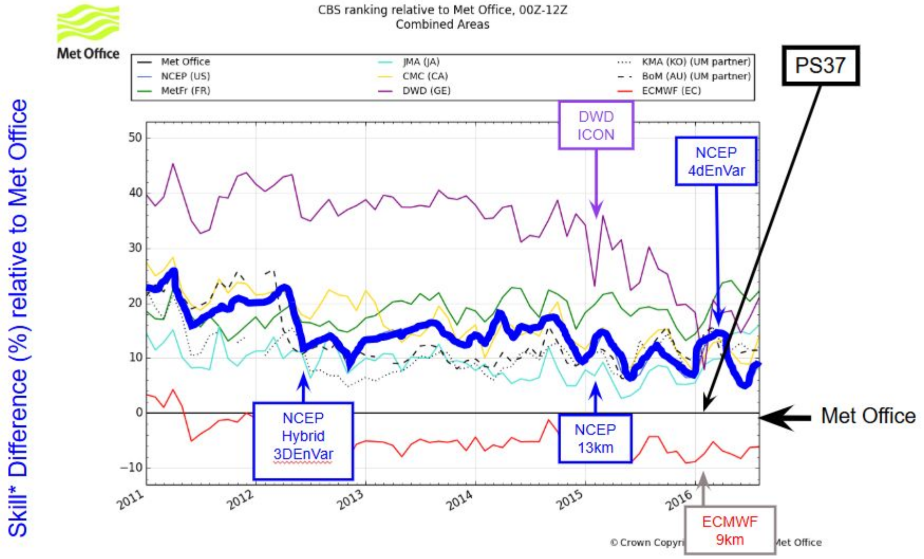
- **Data assimilation must be fast, but also reliable and reproducible**
- **Reliable:**
  1. Run every cycle as planned without crashes: no failures due to overflow/underflow, memory, etc.
  2. Cope with no (or bad) data
  3. Reliable completion time
- **Reproducible:**
  1. Order of operations can change results – big issue for managing large processor count (MPI) and global computations

Computational Capacity Matters

Date	PS	Major Change	KPI impact (obs)
Jul 2011	PS27	DA: Hybrid 4DVar implementation	+2.44 (#1)
Jan 2012	PS28	Minor global changes	-0.40
Mar 2012	PS29	Minor global changes	~0.0
Sep 2012	PS30	Technical change	~0.0
Jan 2013	PS31	DA: Improved hybrid 4DVar	+1.18 (#3)
Apr 2013	PS32	SA and Dust Assimilation	+0.08
Feb 2014	PS33	Rose Implementation	-0.10
Jul 2014	PS34	ENDGAME dynamical core	-0.58
Feb 2015	PS35	SA package + Dust	+0.86 (#5)
Aug 2015	PS36	Transition from IBM to Cray	+0.68
Mar 2016	PS37	DA/SA: VarBC/NewSats/CVT	+2.40 (#2)
Nov 2016	PS38	SA package	+0.99 (#4)
July 2017	PS39	17->10km UM	+0.53
Nov 2017	*PS40*	*DA: CVT/waveband localisation (to date)*	*+0.75*

Top 5 Global NWP Impacts due to improvements in initial conditions

Source: Dale Barker, UK Met Office



* Parameters: Surface pressure, 500hPa geopotential height, 250hPa/850hPa Winds; Forecast ranges from T+24h to T+120h

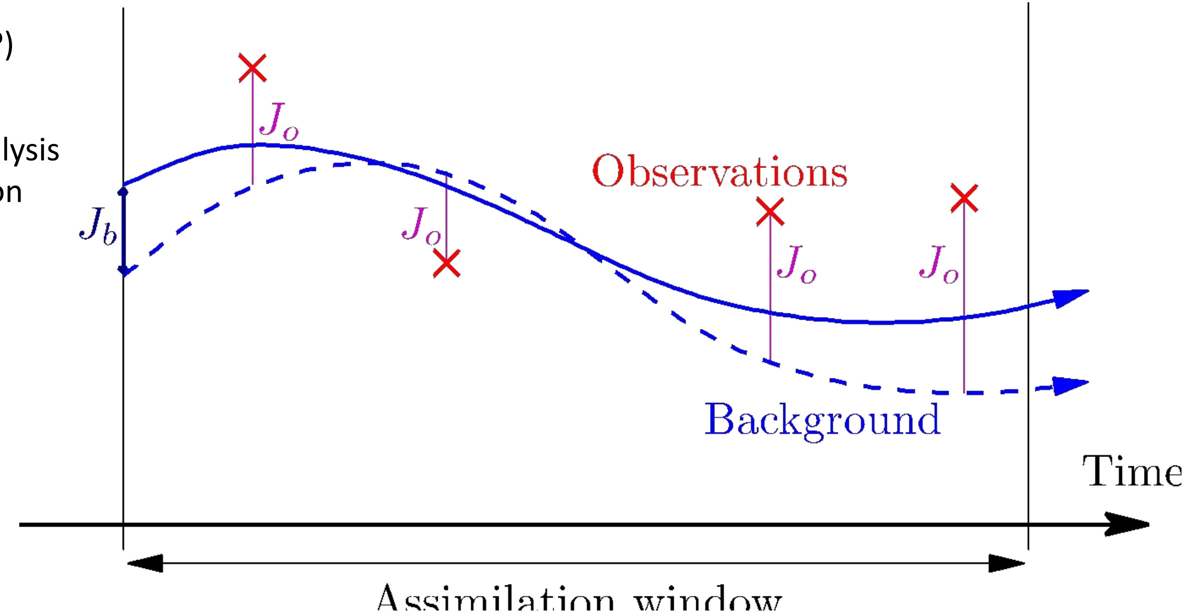
Variational DA

Courtesy JCSDA

$$J(x) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_b) + \frac{1}{2}[\mathcal{H}(\mathbf{x}) - \mathbf{y}]^T \mathbf{R}^{-1}[\mathcal{H}(\mathbf{x}) - \mathbf{y}]$$

Variational Data Assimilation is used by most (all?) operational centers (GSI, NAVDAS, IFS, VAR, ...)

Principle: minimize the distance between the analysis and all available observations over the assimilation window



- $\mathbf{x}_b$ Background state
- $\mathbf{y}$ Observations
- $\mathcal{H}$ Observation operator
- $\mathbf{B}$ Background error covariance
- $\mathbf{R}$ Observation error covariance

Variational DA

Courtesy JCSDA

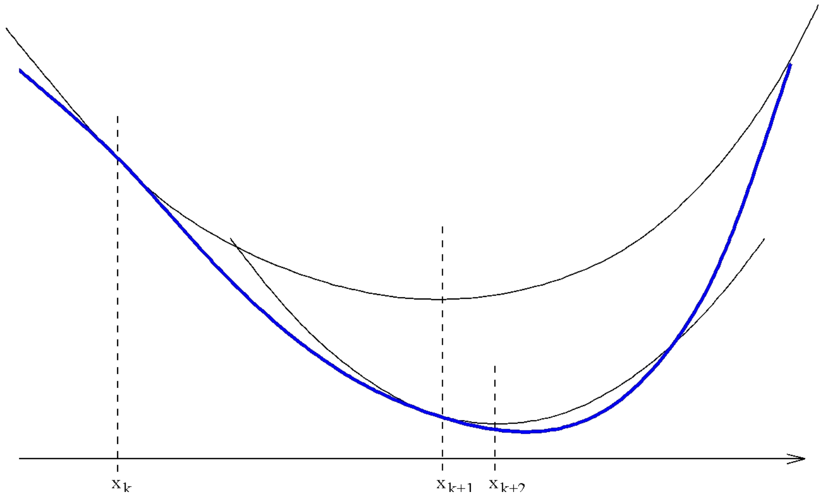
$$J(x) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_b) + \frac{1}{2}[\mathcal{H}(\mathbf{x}) - \mathbf{y}]^T \mathbf{R}^{-1} [\mathcal{H}(\mathbf{x}) - \mathbf{y}]$$

Computational issue: The nonlinear cost function is minimized iteratively

Gauss-Newton outer loop

Nonlinear cost function approximated by series of quadratic problems

Only one operational center (ECMWF) performs more than 1 outer iteration, some perform a partial re-linearization (*middle* loop).



Variational DA

Courtesy JCSDA

Setting the gradient of the quadratic cost function to zero leads to the linear system:

$$\delta\mathbf{x} = (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{d}$$

Computational issue: the size of the data assimilation problem

- The size of **x** is $O(10^9)$
- The size of **y** is $O(10^7)$ and growing
- The observation error covariance matrix is diagonal (or nearly)
- The background error covariance matrix cannot be stored: it must be modeled and coded as a series of linear operators
  - Spectral, Wavelets, Recursive filters, Diffusion operator…
- Even vectors (**x** and **y**) have to be distributed across many processors to fit in memory
  - Adds complexity, especially for non local observations

Variational DA

Courtesy JCSDA

$$\delta\mathbf{x} = (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{d}$$

Computational issue: the covariance matrices need inverting

- The observation error covariance matrix is diagonal (or nearly), for both computational and scientific reasons
- The background error covariance matrix is very ill-conditioned
  - Inversion would be too expensive
  - Change of variable so that $\mathbf{B}^{-1}$ is not needed, for example $\chi = \mathbf{B}^{-1/2}\delta\mathbf{x}$ leads to the linear system:

$$\chi = (\mathbf{I} + \mathbf{B}^{T/2}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{B}^{1/2})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{d}$$

Note: using $\mathbf{B}^{-1}\delta\mathbf{x}$ and $\delta\mathbf{x}$ is another popular choice

Variational DA

Courtesy JCSDA

$$\chi = (\mathbf{I} + \mathbf{B}^{T/2}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{B}^{1/2})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{d}$$

- The linear system is solved by another iterative algorithm: inner loop
  - Conjugate gradient
  - Quasi-Newton

- *Computational issue*: repeated computation of **H** and $\mathbf{H}^T$ is still too expensive
  - The inner loops are run at reduced resolution

- *Computational issue*: preconditioning is essential
  - 25-50 iterations to solve a problem of size $O(10^8 \times 10^8)$
  - NCEP (4DEnVar) performs 50 x 150 iterations

Cost Breakdown (Example)

Courtesy JCSDA

	Time (sec.)	%
$\mathcal{M}$	500	19.5
M	1143	44.5
$\mathcal{H}$	71	2.8
H	228	8.9
B	120	4.7
R	59	2.3
Lin. algebra	53	2.0
I/O	≈305	≈11.8
Other	92	3.6

IFS 4D-Var (12h window)

T1279/T255/T319/T399

≈30M active observations

528 MPI tasks, 18 OpenMP threads

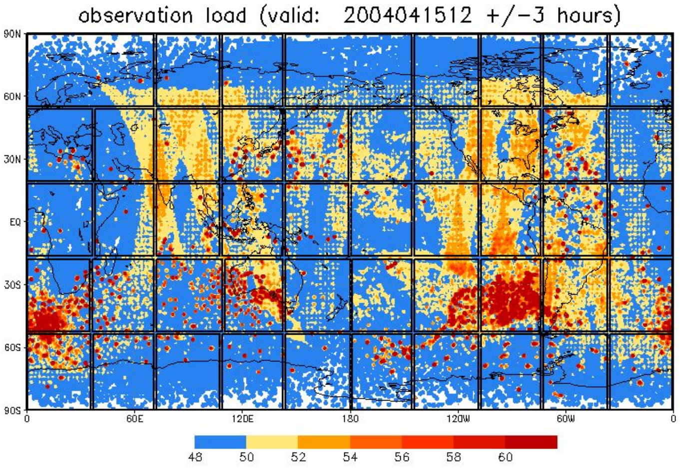
264 nodes, 6336 cores

≈43 min. wall clock time

3D-Var (FGAT) would take half the time (no M and only 1 $\mathcal{M}$ instead of 4)

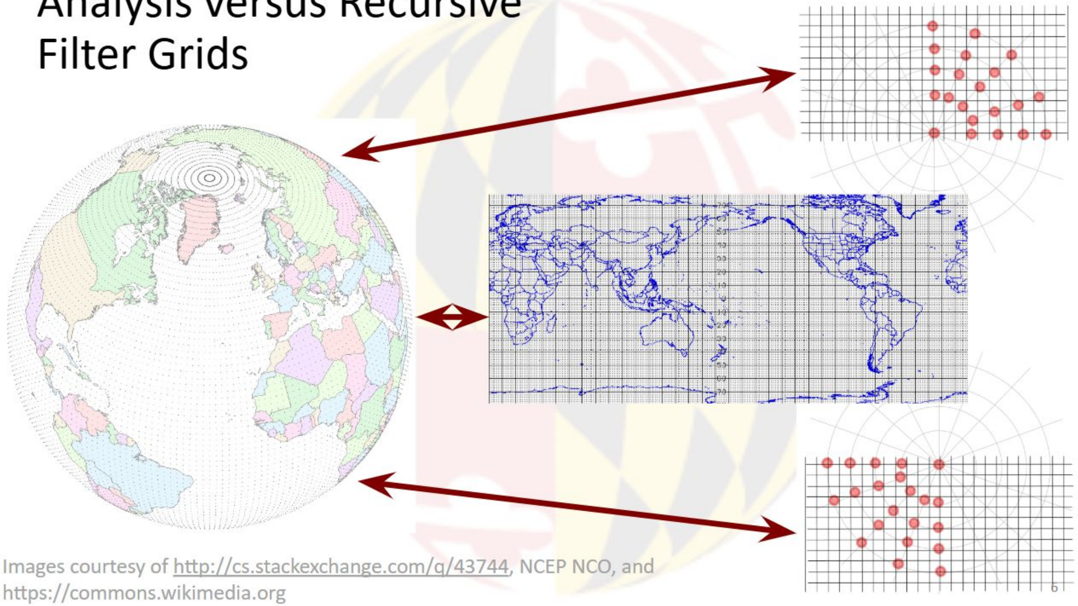
Dual (observation) space algorithms are possible but the gain is small

GSI Example



- The main domain decomposition in GSI is done on evenly sized subdomains
  - All variables/levels stored
  - Schematic example for 50 MPI tasks

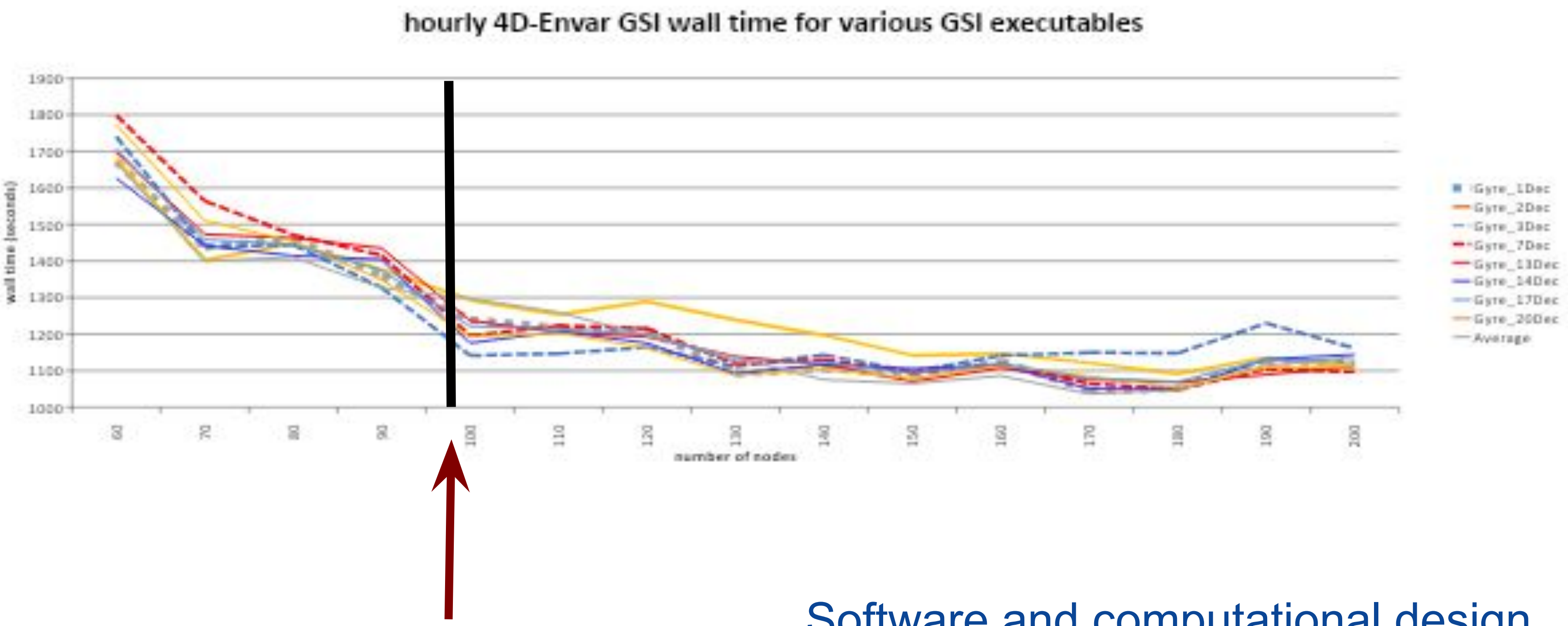
Analysis versus Recursive Filter Grids



Images courtesy of http://cs.stackexchange.com/q/43744, NCEP NCO, and https://commons.wikimedia.org

Background error correlations modeled with Recursive Filters. Above shows domains used. Lots of data movement.

GSI Example



97 nodes = 388 MPI tasks in these runs

Software and computational design matters!

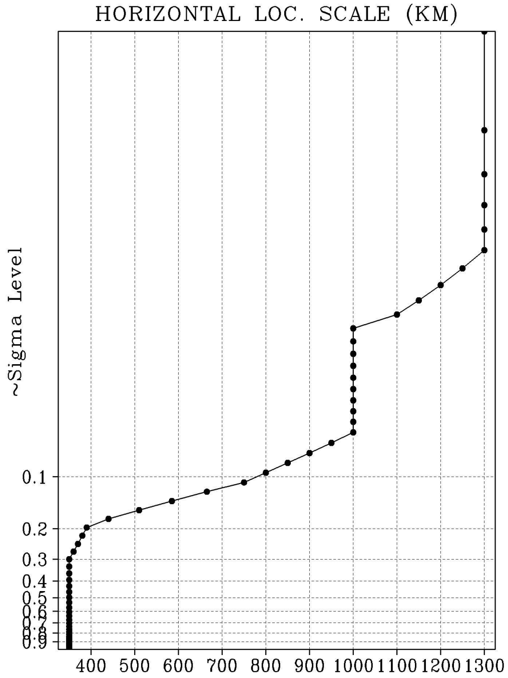
Current Operational GDAS (Hybrid 4DEnVar)

$$J(\mathbf{x}'_c,\mathbf{a}) = \beta_c \frac{1}{2}(\mathbf{x}'_c)^T \mathbf{B}_c^{-1}(\mathbf{x}'_c) + \beta_e \frac{1}{2}\mathbf{a}^T \mathbf{L}^{-1}\mathbf{a} + \frac{1}{2}\sum_{k=1}^{K}(\mathbf{H}_k \mathbf{x}'_{(t)k} - \mathbf{y}'_k)^T \mathbf{R}_k^{-1}(\mathbf{H}_k \mathbf{x}'_{(t)k} - \mathbf{y}')$$

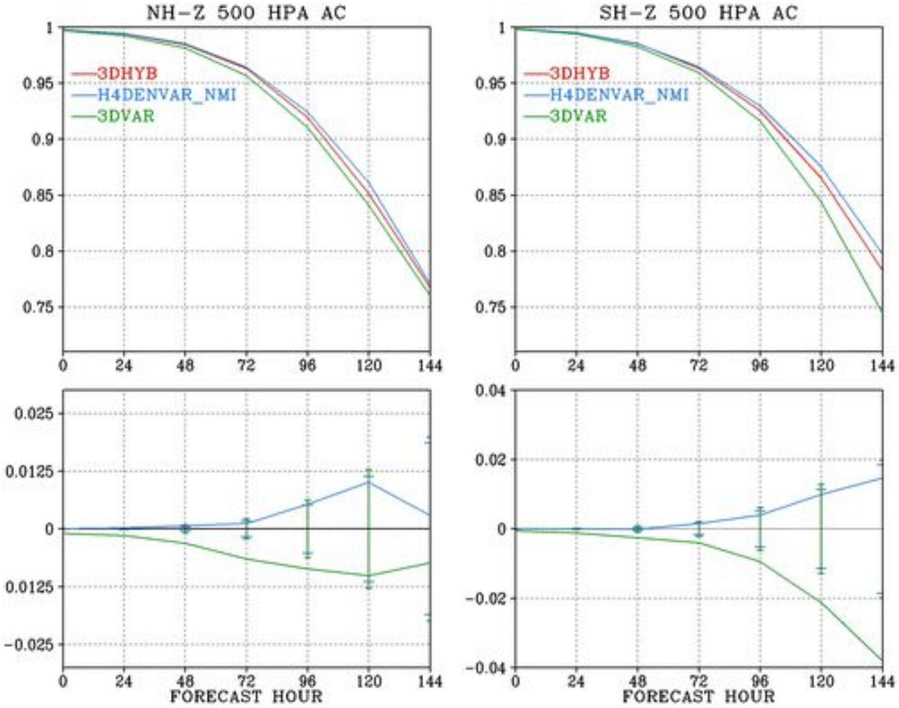
$$\mathbf{z} = \mathbf{B}^{-1}\mathbf{x}'_c \qquad \mathbf{v} = \mathbf{L}^{-1}\mathbf{a}$$

C768L64 (~12km) **FV3-based GFS**

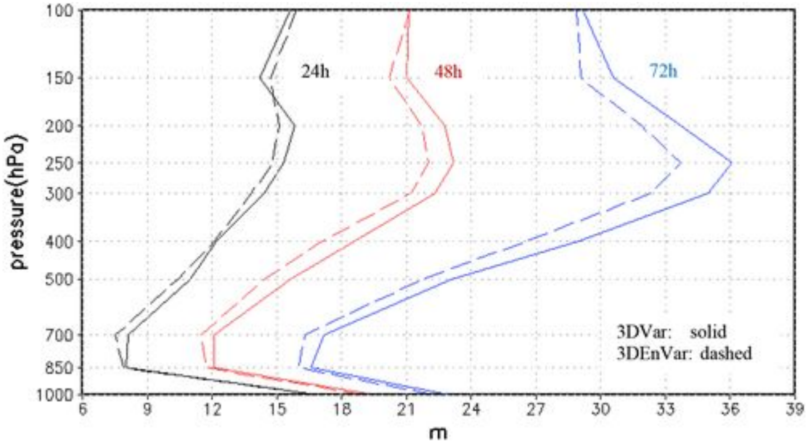
- 80 member ***C384L64 (~25km)*** EnSRF for data assimilation
- Level-dependent localization
- Stochastic physics to represent model uncertainty (SPPT, ~~SKEB~~, SHUM) – Since January 2015 + RTPS (no more additive perturbations)
- Analysis increment at ensemble resolution
- Ensemble perturbations centered about hybrid analysis
  - Ensemble mean state estimate replaced



Ensemble Importance for Forecast Skill



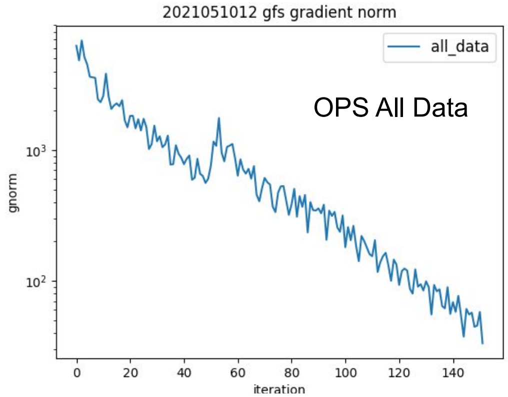
Kleist and Ide (2015), Fig. 10: OSSE based study. Comparison of skill from 3DVar, Hyb 3DEnVar, and Hyb 4DEnVar for simulated August 2005.



Wu et al. (2017), Fig. 10: Impact of using global ensemble perturbations in regional hybrid EnVar (forecast fits to geop. height, radiosondes). This has also been realized in other regional systems at NOAA

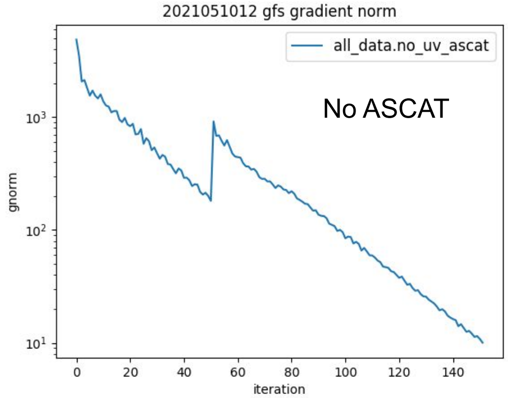
- **Now standard (in some form) at most (all?) NWP centers**
- **Has changed computational distribution**
  1. Running short, ensemble forecasts now most expensive part of DA cycle

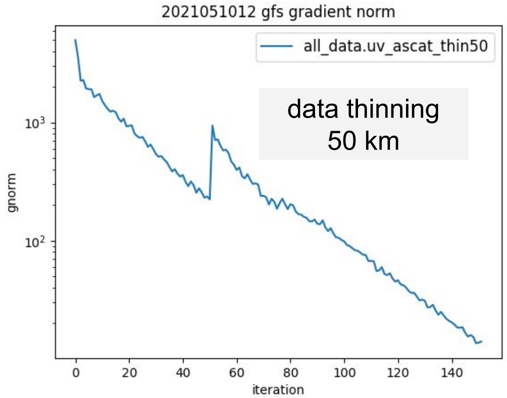
Minimization Issue

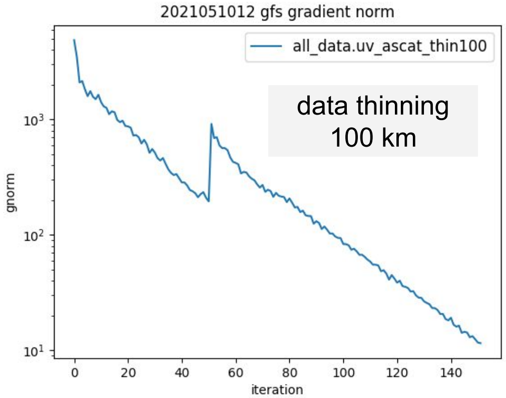


The operational GFS occasionally detects and reports problematic GSI minimizations

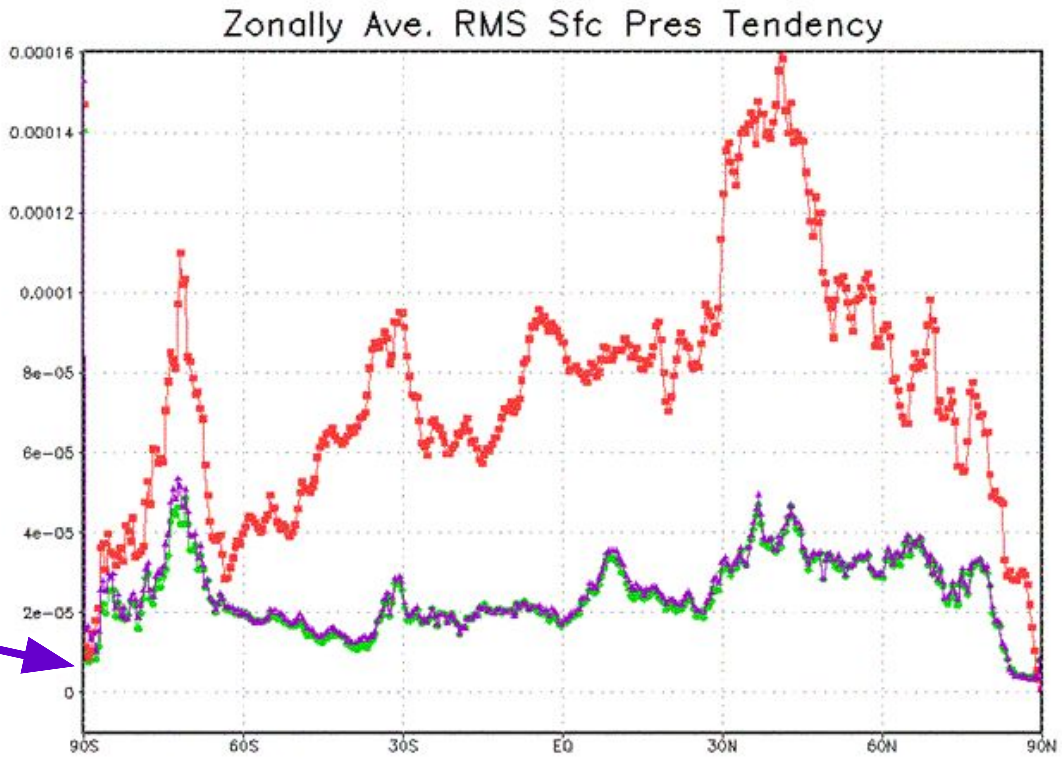
- Example: the operational 2021051012 GFS cycle
- Investigation of this case found that following changes to the use of ASCAT yielded a much smoother minimization:
  - Remove all ASCAT winds
  - Thin ASCAT data with 50 or 100 km grid box







Balance / Noise



Minimal increase with TLNMC

Zonal-average surface pressure tendency for background (green), unconstrained GSI analysis (red), and GSI analysis with TLNMC (purple)

Operational Summary

- **Delicate balance to maximize science through efficiency**
  - Operational drivers often require compromise
  - Sometimes, the best ideas simply are not feasible from the perspective of operational requirements
- **Observations:**
  - They always change and job is never done
  - What would you do if commercial aircraft stopped flying (2020 pandemic)?
  - What would you do if we suddenly lost lots of rawinsondes?
- **The future is bright:**
  - Progress in AI/ML brings unique opportunity to reduce/remove some of the current pain points

NWS/NCEP/EMC Development Strategy

- Introduction
- Advanced Infrastructure / JEDI
- Research and Development
  - Use of Observations
  - New and Upcoming Observations
  - Algorithms
  - Toward Continuous DA
  - Coupled DA
  - AI/ML
  - Reanalysis
  - Development Practices
- Data Assimilation Vision - Holistic Approach
  - JEDI, Partnerships, Embracing Change, and Workforce
  - Risk Management

Data Assimilation Strategy for NOAA/NWS/NCEP/EMC

U.S. Department of Commerce
National Oceanic and Atmospheric Administration
National Weather Service
National Centers for Environmental Prediction
5830 University Research Court
College Park, MD 20740-3818

Office Note 517
https://doi.org/10.25923/kqra-6z66

Data Assimilation Strategy and Development Plan for NCEP's Environmental Modeling Center

Daryl Kleist, Jacob R. Carley, Andrew Collard, Emily Liu, Shun Liu, Cory R. Martin, Catherine Thomas, Russ Treadon, Guillaume Vernieres

NOAA/NWS/NCEP Environmental Modeling Center

College Park, Maryland

January 2024

https://doi.org/10.25923/kqra-6z66

AI/ML for Operational DA Priorities

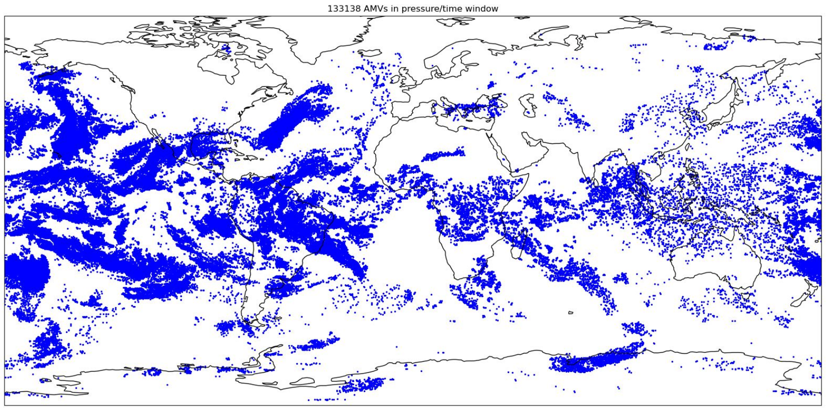
From EMC DA Strategy, Section 3.6 Summary Table

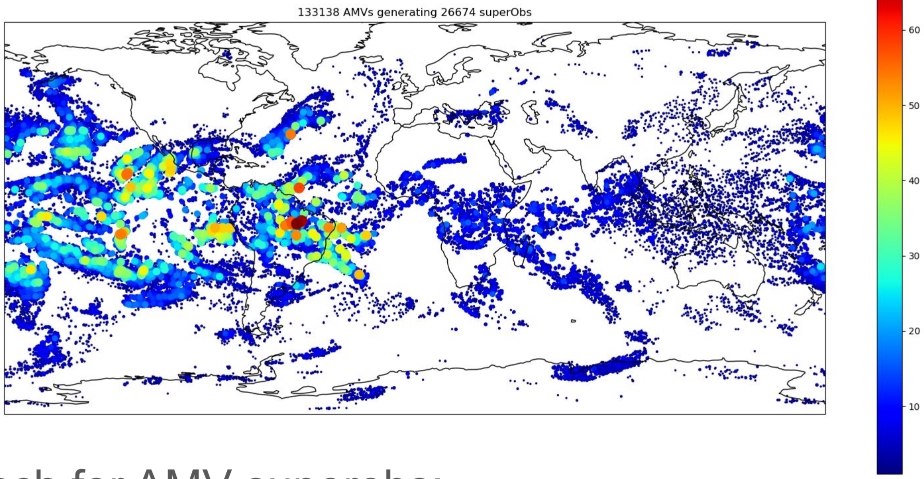
1. **Observations –** quality control, data selection, bias correction, superobservations, extraction of maximal information content, anomaly detection and operational monitoring;
2. **Forward operator emulation** – computational efficiencies, replacement for complex operators;
3. **Background error** – computational efficiencies, multivariate aspects and coupled assimilation, parameter estimation for error models;
4. **Background** – dynamic downscaling, bias correction;
5. **Model error** – estimation and correction;
6. **Emulator exploitation** – replacement for TL/AD in 4DVar, efficient creation of huge ensembles to avoid localization;
7. **Hybridization** – explicit blending of ML and DA; joint frameworks; pathways to going directly from observations to simulation/emulation.

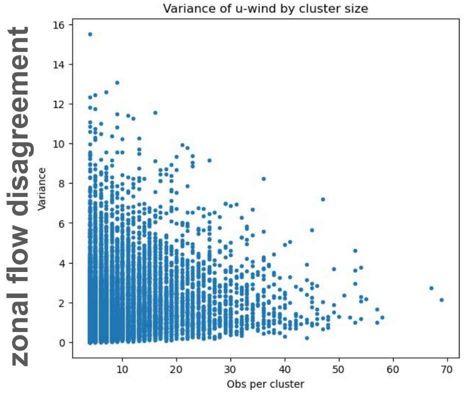
EMC DA Team now has an AI/ML “Study/Working Group” – pursuing many new AI-based projects both small and large

Atmospheric Motion Vectors: Clustering for Superobservations

133,000 AMVs ------- clustering --> 27,000 super observations



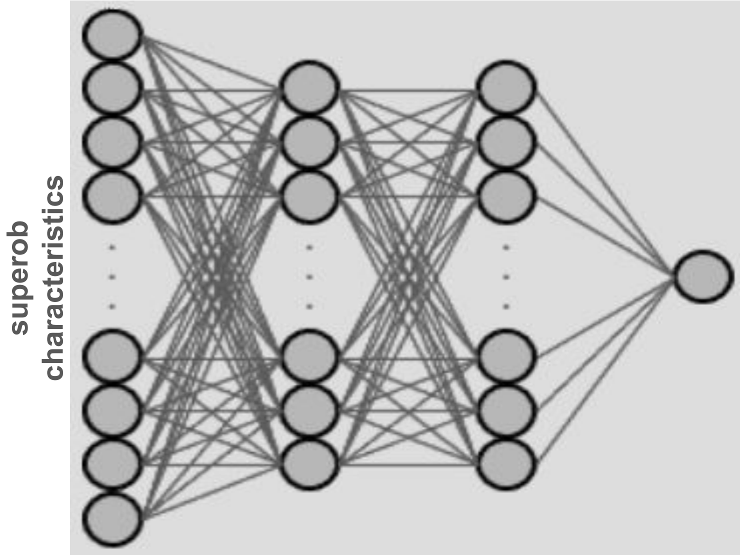




A clustering approach for AMV superobs:

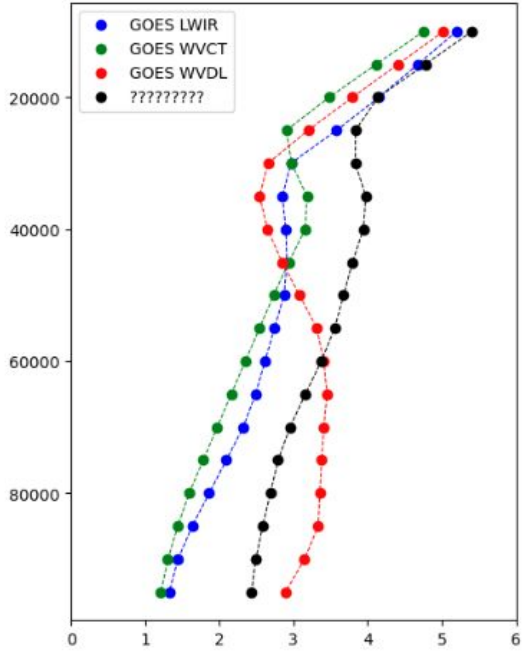
- reducing data volume (operational efficiency)
- combining similar, nearby AMVs (focus on information content)
- addressing spatial error correlation (alternative to thinning)
- producing unified set of observations from disparate AMV sources

Atmospheric Motion Vectors: Superobservation Error Estimation



fully connected, deep learning neural network

superob error estimate



- Error estimation/assignment is critical for assimilation
- Ability to learn type-dependency on error; differences among GOES types is similar to observed in departure statistics
- With same profile but no ob type information, can predict generic error profile (upper right, black) that appears reasonable

Toward ML-based Quality Control (example: aircraft)

typ	tot	pqc<7	pqc>7	pqc=8	pqc=9	pqc=10	pqc=11	pqc=12	pqc=13	pqc=14	pqc=15	use=-1
230	1197|	1128|	69|	0	0	0	0	0	69	0	0	83
231	18524|	16892|	1632|	0	0	0	0	0	1632	0	0	1899
233	144779|	138095|	6684|	0	0	0	0	0	6684	0	0	6778
234	2300|	1753|	547|	0	0	0	0	0	547	0	0	579
235	522|	516|	6|	0	0	0	0	0	6	0	0	6
236	338|	338|	0|	0	0	0	0	0	0	0	0	0
tot	167660|	158722|	8938|	0	0	0	0	0	8938	0	0	9345

Observation Inventory For Quality Control Training by Type

TRAIN THE AI/ML/QC MODEL

read

```
# import the training dataset
import pandas as pd
import matplotlib.pyplot as plt
from pandas.api.types import is_string_dtype, is_numeric_dtype
train = "trains/train-air-uv"
preds = "trains/preds-air-uv"
model = "clf-air-uv"
df = pd.read_csv(train)
df.head()
```

prepare

```
# prepare X and y for training
from sklearn.model_selection import train_test_split
from sklearn import preprocessing
X = df.drop(columns=['qmk','pmk'])
y = df['qmk']

# scale the training data to unit variance
from sklearn import preprocessing
scaler = preprocessing.StandardScaler()
X = scaler.fit_transform(X)
```

train

```
# fit and run the model on trining data
from sklearn.neural_network import MLPClassifier
from sklearn.datasets import make_classification
clf = MLPClassifier(hidden_layer_sizes=(50,50),random_state=1, max_iter=200, verbose=True)
clf.fit(X,y); yp=clf.predict(X)
```

display

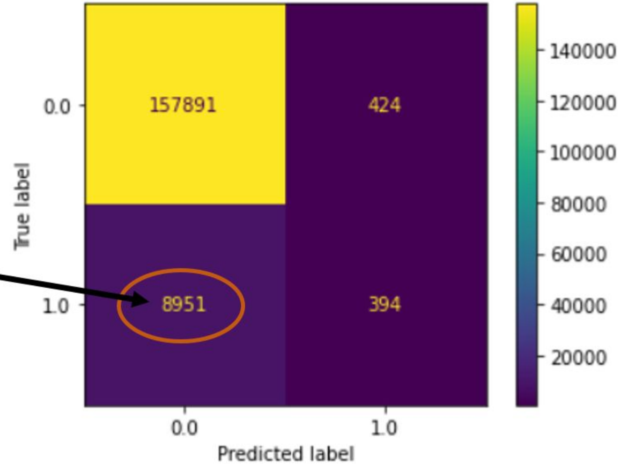
```
# show the confusion matrix
from sklearn.metrics import ConfusionMatrixDisplay
ConfusionMatrixDisplay.from_predictions(y,yp)
plt.show()
```

save

```
# save the trained model file
import joblib
joblib.dump(clf, model)
```

Attempting to learn Quality Control for aircraft based observations

Need to examine why the AI wants to accept so many observations rejected by ACQC/GSI



	typ	lat	lon	prs	umf	vmf	qmk	pmk	prd
0	233.0	-10.997	-0.007	227.3	3.921	7.448	0.0	1.0	0.0
1	231.0	-5.400	8.927	227.4	0.031	-1.379	0.0	1.0	0.0
2	231.0	-4.743	8.283	227.0	-0.699	-5.167	0.0	1.0	0.0
3	231.0	-4.080	7.637	227.4	-4.802	-2.532	0.0	1.0	0.0
4	231.0	-3.413	6.987	227.4	-2.920	-1.777	0.0	1.0	0.0

training target metadata prediction

Bias Correction

- Variational Bias Correction has been critical for advanced/improved use of satellite observations (extended to other observation types)
  - Extremely effective - deals with instrument drift, representativeness issues, etc.
  - However, makes observations “look” like background in absence of anchor observations

Linear predictor model for bias in each channel:

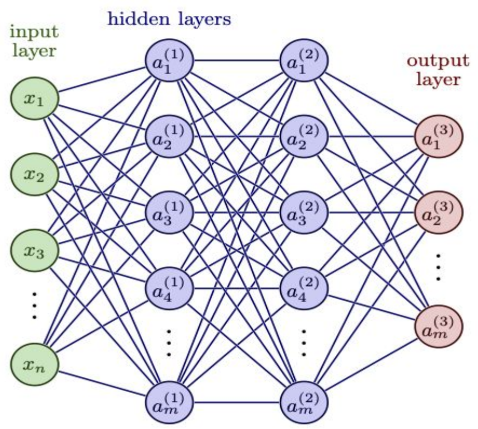
$$\mathbf{b}(\mathbf{x}, \beta) = \sum_{i=0}^{N_p} \beta_i \mathbf{p}_i(\mathbf{x})$$

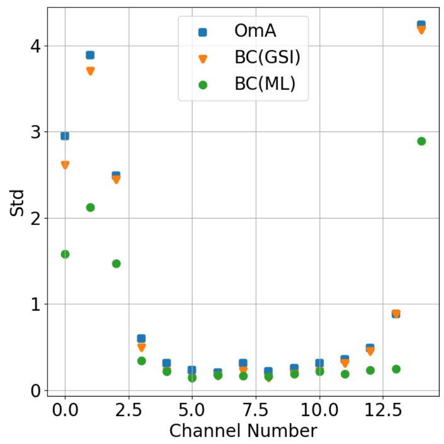
Cost function:

$$J(\mathbf{x}, \boldsymbol{\beta}) = \overbrace{(\mathbf{x}_b - \mathbf{x})^T \mathbf{B}_x^{-1} (\mathbf{x}_b - \mathbf{x})}^{J_b\text{: background constraint for x}} + \overbrace{(\boldsymbol{\beta}_b - \boldsymbol{\beta})^T \mathbf{B}_\beta^{-1} (\boldsymbol{\beta}_b - \boldsymbol{\beta})}^{J_\beta\text{: background constraint for } \beta}$$
$$+ \underbrace{[\mathbf{y} - \mathbf{b}(\mathbf{x}, \beta) - h(\mathbf{x})]^T \mathbf{R}^{-1} [\mathbf{y} - \mathbf{b}(\mathbf{x}, \beta) - h(\mathbf{x})]}_{J_o\text{: bias-corrected observation constraint}}$$

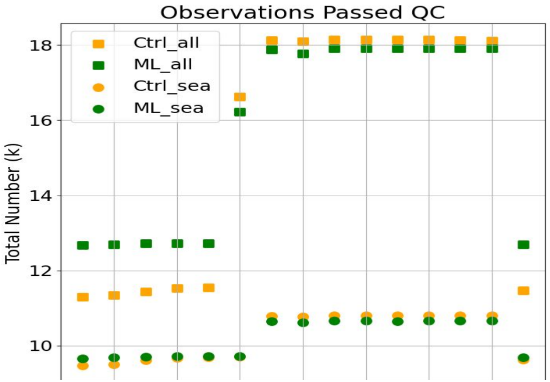
ML-based Bias Correction

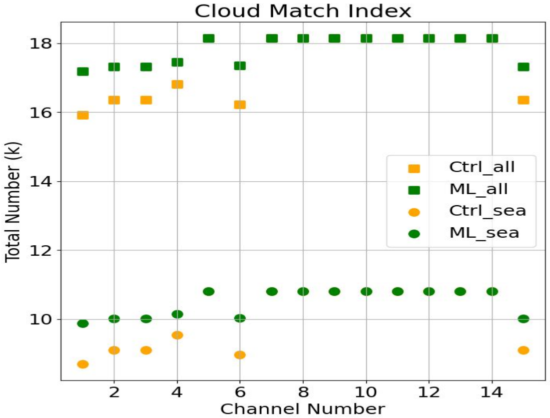
- **Build prototype ML model for bias correction of AMSU-Amp**
  - From Metop-C, 2023: 70% for training; 30% for validation
- **Simple neural network with one input layer, two hidden layers, and one output layer**
  - 53 inputs (based on GSI predictors) 50 neurons for each hidden layer, 15 outputs to bias correction for 15 channels
- **Summary of initial attempts:**
  - Number of observations passing quality control has increased
  - Error model produces smaller errors
  - Ability to assimilate more cloud-impacted pixels



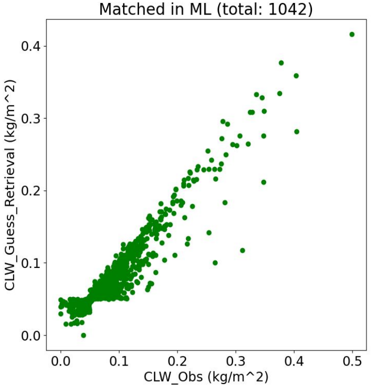


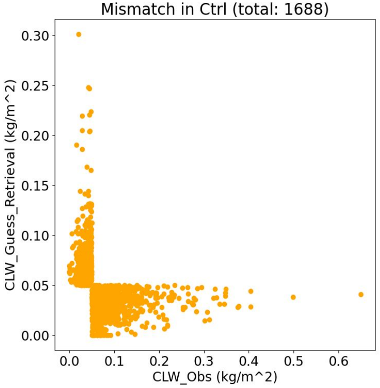
ML-based Bias Correction



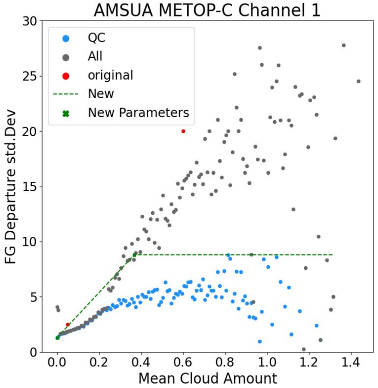


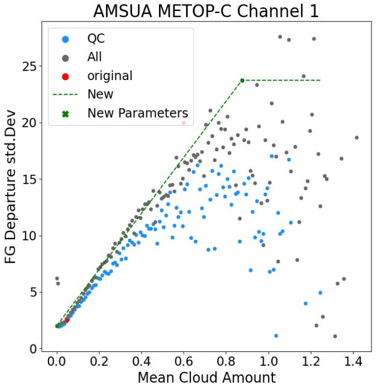
Match in ML .vs. Mismatch in Ctrl





Small Error (ML) .vs. Large (Ctrl)





Summary

- Expansion and acceleration of AI work is one of the pillars of EMC DA Strategy
- Study/Working group rapidly progressing on a variety of AI/ML projects for DA
- Many early successes with operational pathways (improve efficiency and skill in DA/prediction)
- Early work underway toward even more radical uses: e.g. direct prediction from observations (similar to GraphDOP work at ECMWF)

?

Questions

daryl.kleist@noaa.gov