Impacts of machine learning for selective data thinning in data assimilation

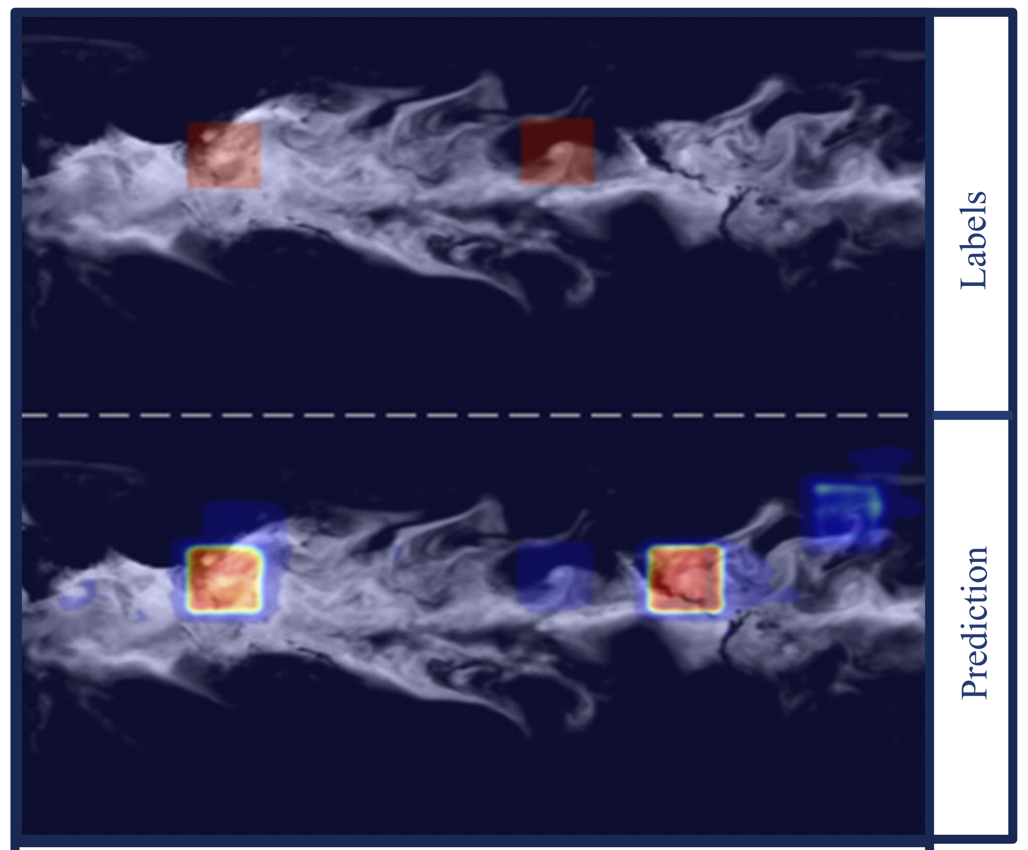
Kirana Bergstrom[1,2], Christina Kumler[1,2], Isidora Jankov[2], Jebb Q. Stewart[2]

1: Cooperative Institute for Environmental Sciences at University of Colorado Boulder, 2: NOAA Global Systems Laboratory

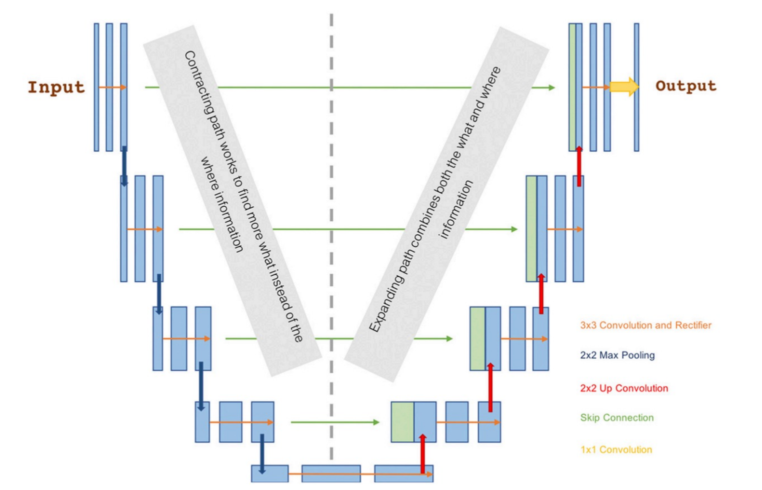
Workflow, software tools and methods for intelligent data thinning in data assimilation

Machine learning for cyclone identification

Previously existing cyclone ML algorithm ([1]) produces images with each pixel labeled with a value in [0,1] representing confidence of "cyclone" or "no cyclone". A threshold is set for binary classification of each pixel.



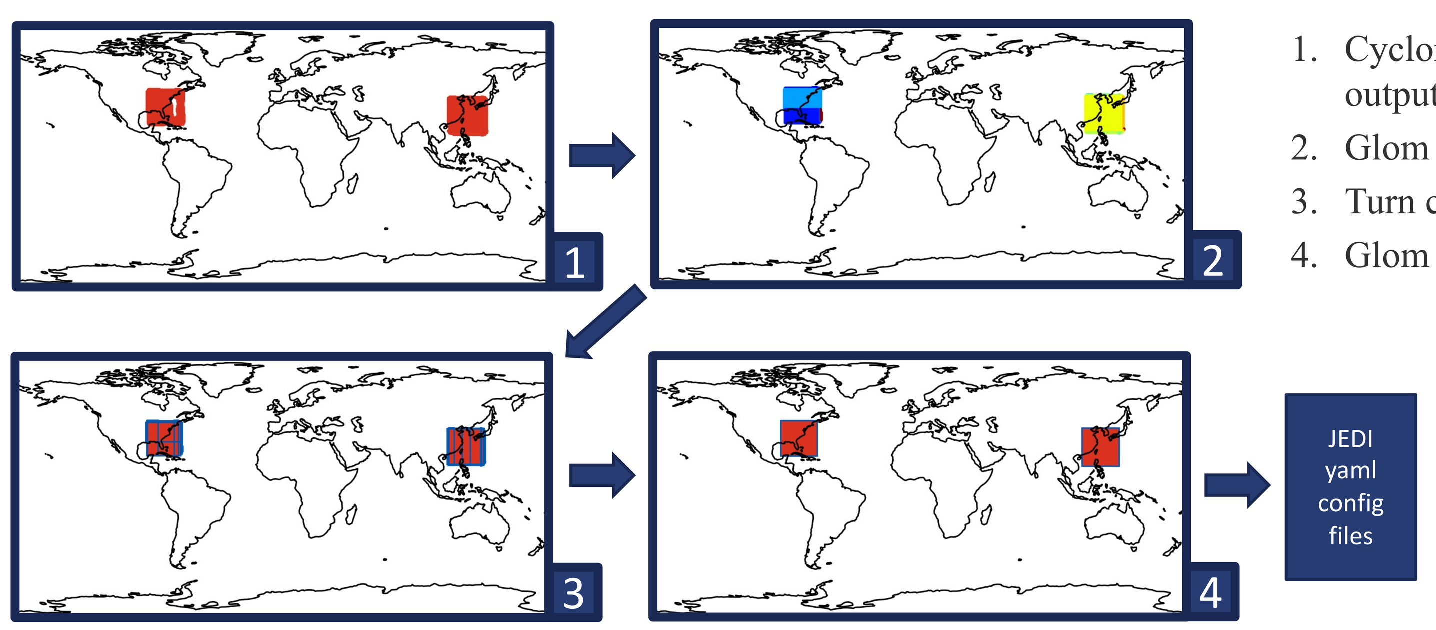
2017-08-23:12:00:00



The cyclone ML algorithm is a convolutional neural network (CNN) with a U-Net architecture (shown above). Cyclone labels are from the International Best Track Archive for Climate Stewardship (IBTrACS) tropical cyclone database and predictors are images representing the GFS total precipitable water output data field from three separate timesteps (current time, -6hrs, +6hrs).

Glue code: ML output to DA input

We use a piece of "glue" code to translate the image output from the cyclone ML model into box-shaped regions that can be used to create geographic observation filters and input into the data assimilation system.



1. Cyclone identification ML algorithm output for threshold = 0.7
2. Glom together cyclone-positive pixels
3. Turn clumps of pixels into boxes
4. Glom together boxes

Result is a set of `yaml` config files that can be input directly into JEDI that are created from a default `yaml` file and covers the add_back configuration.

Data assimilation in JEDI

The Joint Effort for Data assimilation Integration (JEDI) ([2]) is a comprehensive software tool for data assimilation (DA) in earth systems that is under active development. For this project we used JEDI-FV3 for the variational DA application, as well as several specific useful components of JEDI: assimilation of satellite data, observation converters for converting bufr files to IODA files, cold start → warm start, analysis → model, and GFS → lat/lon grid converters, and state difference creators.

The DA specifications that apply for all cases and configurations: we use 3D-EnVar using 30 EnKF GDAS ensemble 6-hr forecasts. The background is a 6-hr UFS forecast from a 0-hr GFS analysis state. The background resolution is C384 and the ensemble and increment resolution is C96 for Case 1, C192 for Case 2. For vertical resolution we use 64 levels.

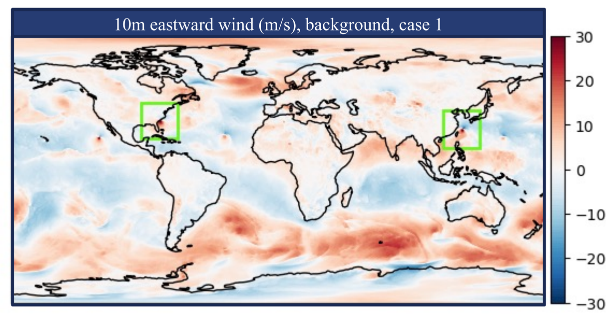
Forecasting in UFS

We use the Unified Forecast System (UFS), a community-based, coupled, comprehensive Earth modeling system, to create the backgrounds for JEDI, to run chgres to generate the ensembles, and to create the forecast from DA analysis results.

Introduction

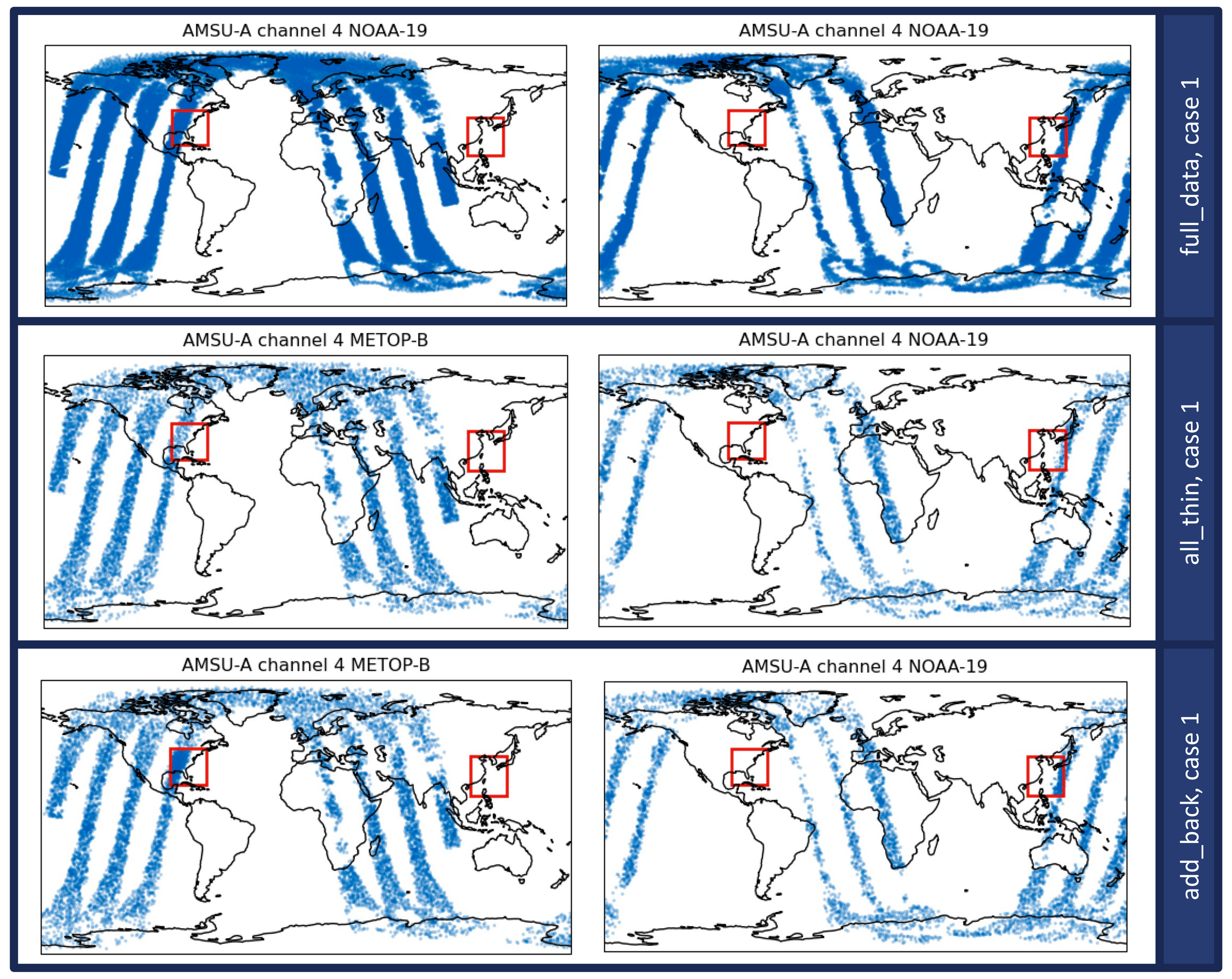
Recent increases in the volume and complexity of environmental information through satellite observations means it has become increasingly challenging to extract and utilize meaningful information in real time from the high density information that is being received. Improvements in the usage of satellite data can result in an improved analysis, and consequently forecast, of NWP models. An existing machine learning (ML) algorithm can identify areas with active and rapidly evolving weather. We will show the process for transforming the ML output, which is given by global image segmentation of pixels with active weather or with no active weather, into the selective filters for thinning that are usable by JEDI, the next generation data assimilation software system. We also present preliminary results for a case study showing that a 3D-EnVar data assimilation algorithm is sensitive to selective thinning of satellite data in the ML-identified regions, and results indicating that global forecasts made from the analysis are also sensitive to this selective thinning.

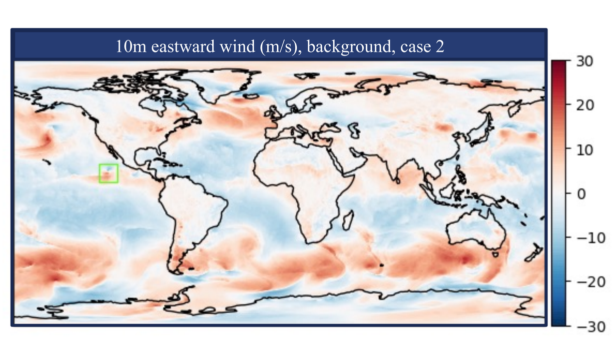
Case studies



Case 1: The first case study is taken on 2019-09-05:18:00:00. This global case includes an cyclone of the South East coast of the United States (Hurricane Dorian) and a cyclone off the South East coast of China that are picked up by the ML model and produce the cyclone zones/regions of interest for this case study.

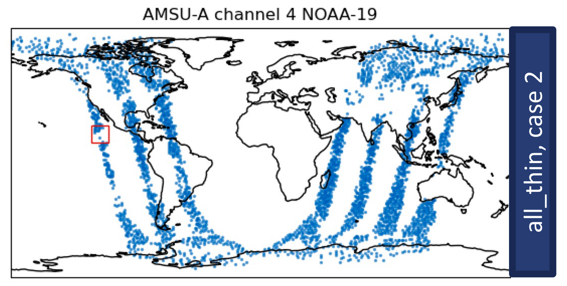
Observations are MHS channels 1-5 and AMSU-A channels 4-6 and 9-14 from satellites NOAA-19 and METOP-B, as well as radiosonde and surface station data. We tested three different configurations: all satellite data (full_data), 75% of all satellite data thinned (all_thin), and 75% of satellite data thinned except in the cyclone region (add_back). All conventional data is used in each case.





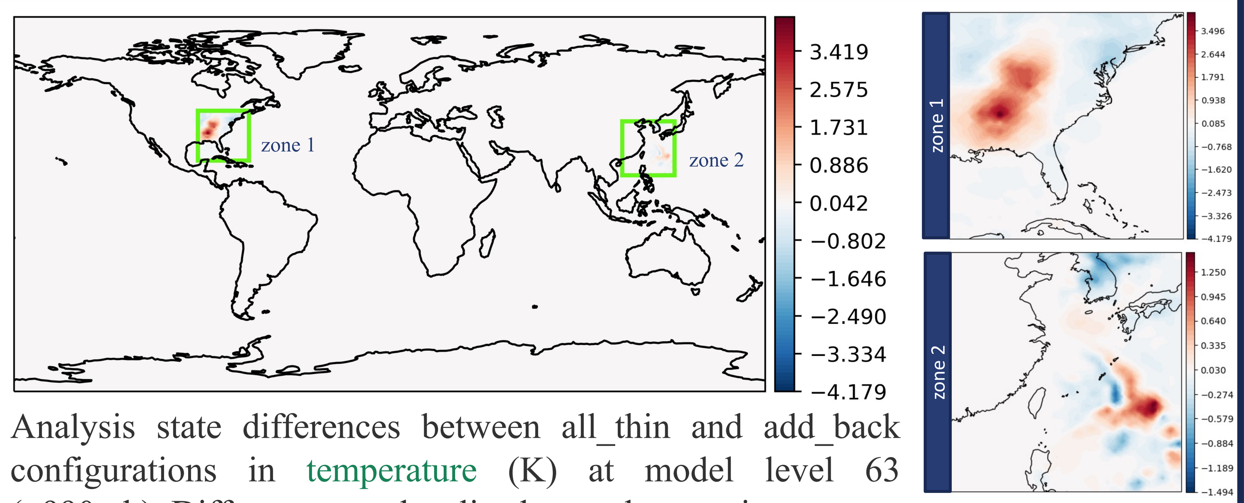
Case 2: The second case study is 2020-10-01:00:00:00 and there is one Pacific cyclone (Hurricane Marie) that produces the cyclone zone/region of interest for this test case.

Observations are AMSU-A channels 4-6 and 9-14 from satellites NOAA-19 as well as aircraft data. We tested three different configurations: all satellite data (full_data), 75% of all satellite data thinned (all_thin), and 75% of satellite data thinned except in the cyclone region (add_back). All conventional (aircraft) data is used in each configuration.



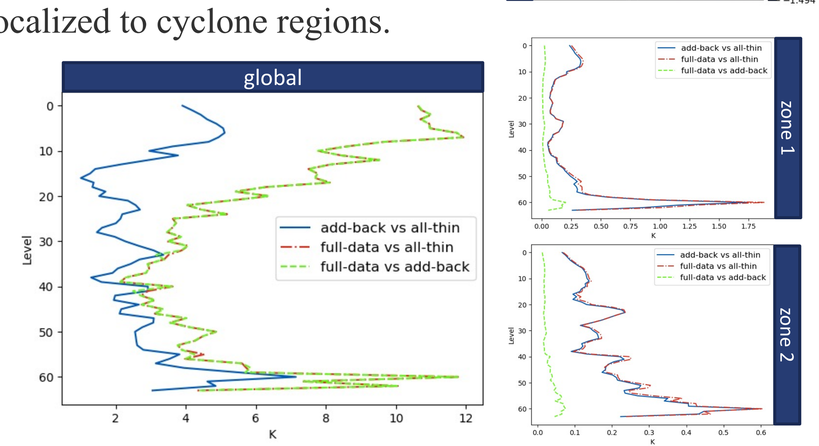
Results

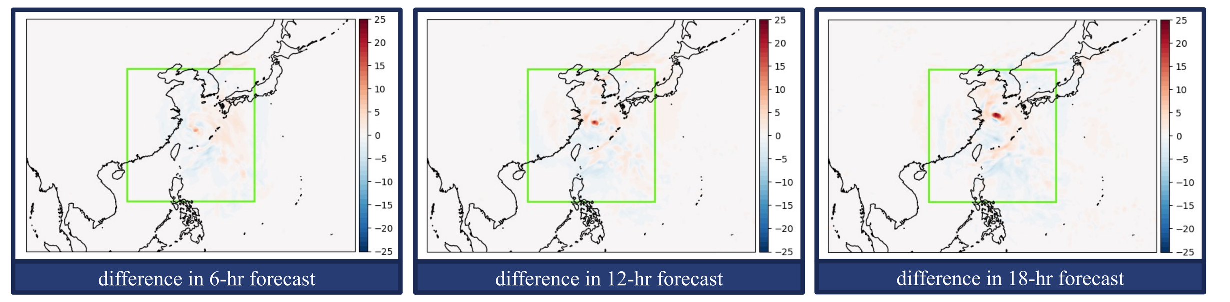
Case 1: 2019-09-05:18:00:00



Analysis state differences between all_thin and add_back configurations in temperature (K) at model level 63 (~990mb). Differences are localized to cyclone regions.

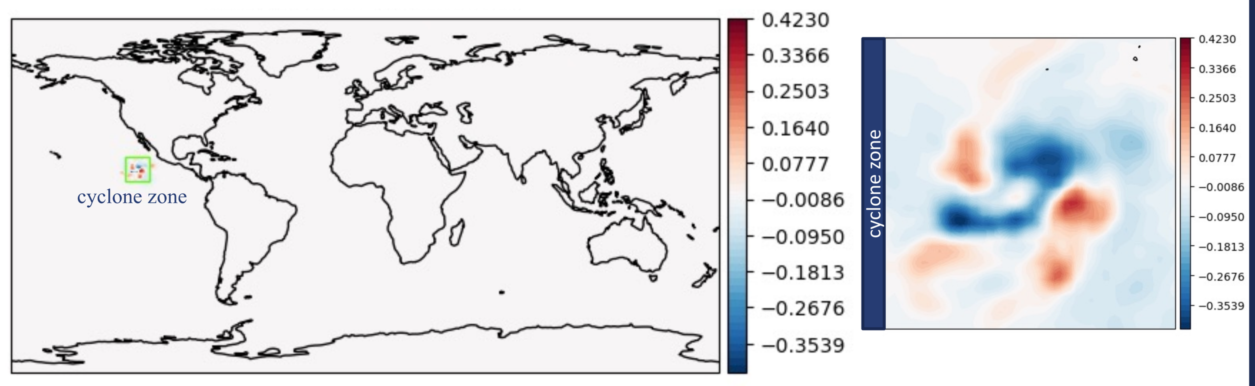
Absolute mean of temperature differences in analysis state in each level globally and in each zone. add_back is close to full_data are in local cyclone zones, but closer to all_thin globally.





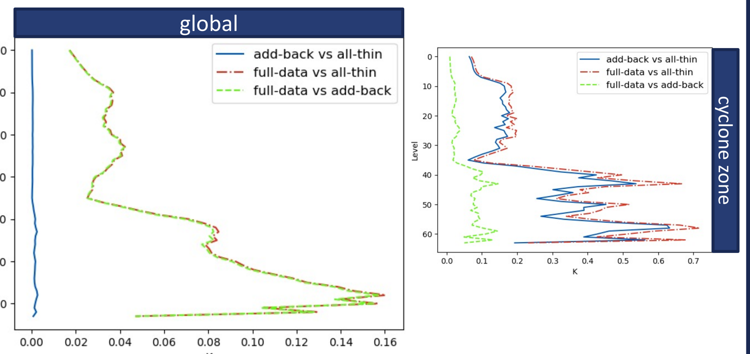
Forecast differences between all_thin and add_back configurations for eastward wind (m/s) at ~850mb over zone 2 showing intensification, spread of the impact of differences in DA analysis as the forecast runs further in time.

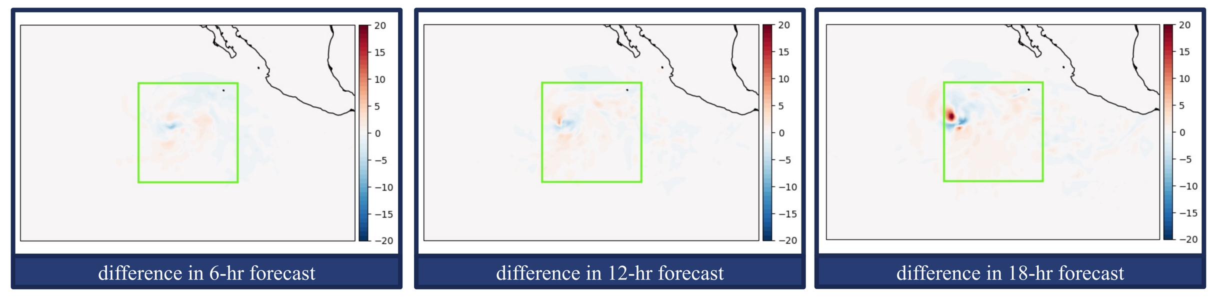
Case 2: 2020-10-01:00:00:00



Analysis state differences between all_thin and add_back configurations in temperature (K) at model level 63 (~990mb). Differences are localized to cyclone zone (smaller than in Case 1 and completely in the satellite's path).

Absolute mean of temperature differences in analysis state in each level globally and in each zone. add_back is close to full_data are in local cyclone zones, but closer to all_thin globally.





Forecast differences between all_thin and add_back configurations for eastward wind (m/s) at ~850mb over zone, showing intensification, spread of the impact of differences in DA analysis as the forecast runs further in time.

Contact

Kirana Bergstrom: kirana.bergstrom@noaa.gov
Christina Kumler: christina.e.kumler@noaa.gov
Isidora Jankov: isidora.jankov@noaa.gov
Jebb Q. Stewart: jebb.q.stewart@noaa.gov

References

[1] C. Kumler-Bonfanti, J. Stewart, D. Hall and M. Govett, "Tropical and extratropical cyclone detection using deep learning", *J. Appl. Meteorol. Climatol.*, vol. 59, no. 12, pp. 1971-1985, Dec. 2020.
[2] UCAR. JEDI. (2017) [Online]. Available: https://github.com/JCSDA/fv3-jedi.
[3] NOAA. UFS. (2012) [Online]. Available: https://github.com/ufs-community/ufs-mrweather-app.