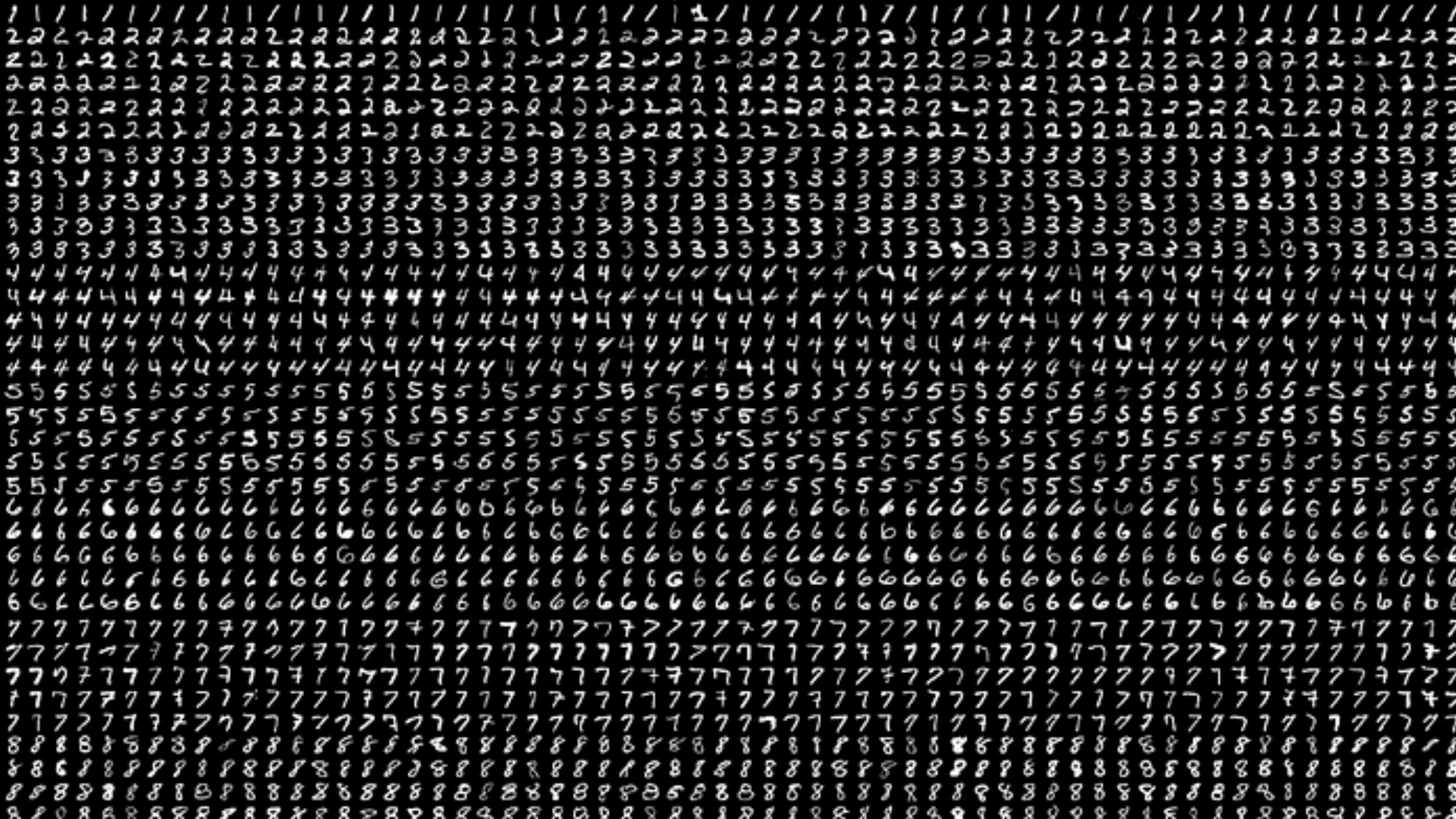


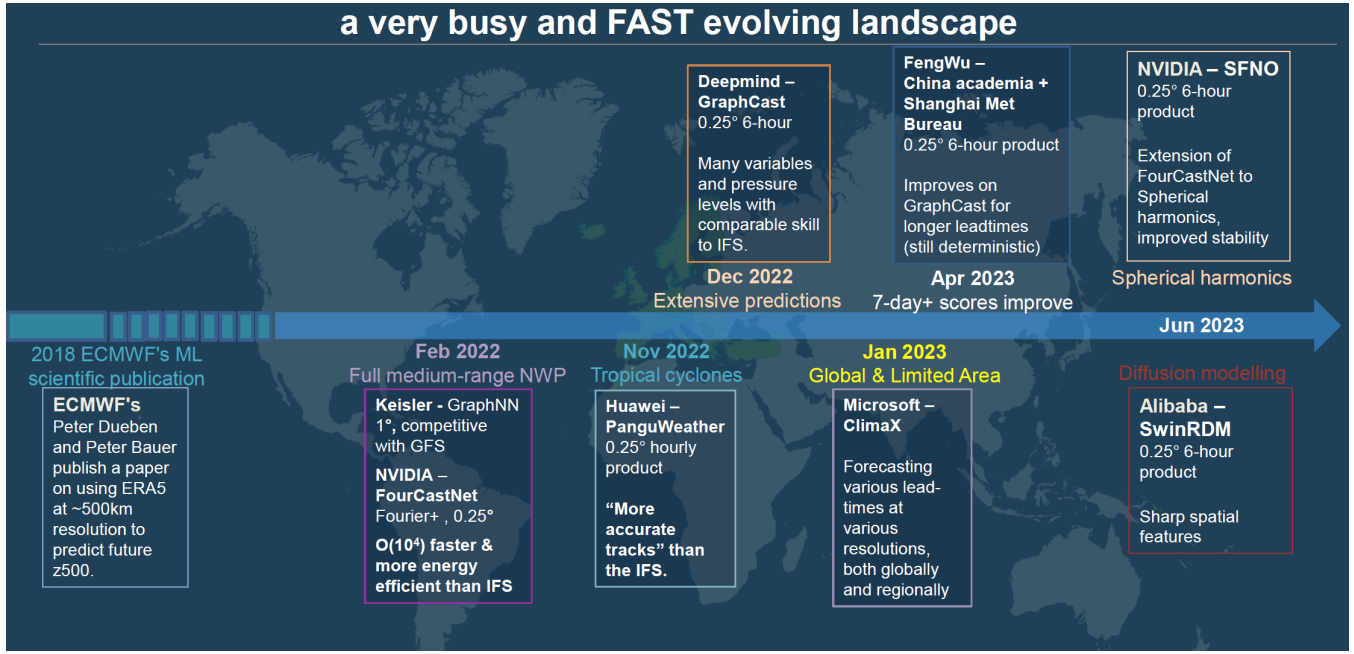
Training datasets available from NOAA

Presenting Sergey Frolov
NOAA/PSL

Credit to datasets goes to many people at NOAA

Presented at: Nov 28, 2023

What makes for a good training dataset?



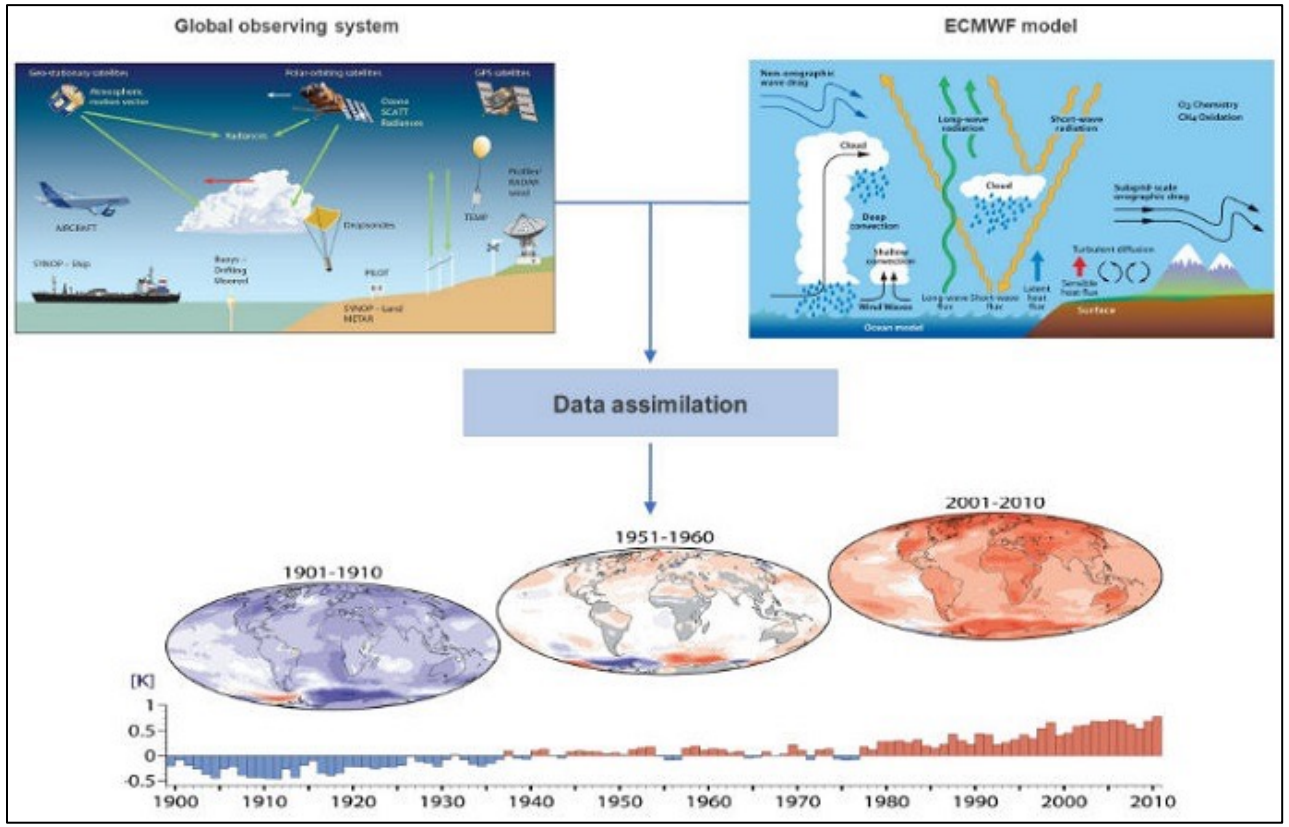
Chantry 2023

- ERA5 has been exclusively used for all emulator training to date.
- What does it mean for the future of NOAA (and NCAR and DOE) models if we don't have comparable training datasets?

Outline

- What makes a good training dataset?
- History of reanalysis datasets at NOAA
- Recommended training datasets
- Future training datasets

What makes for a good training dataset?

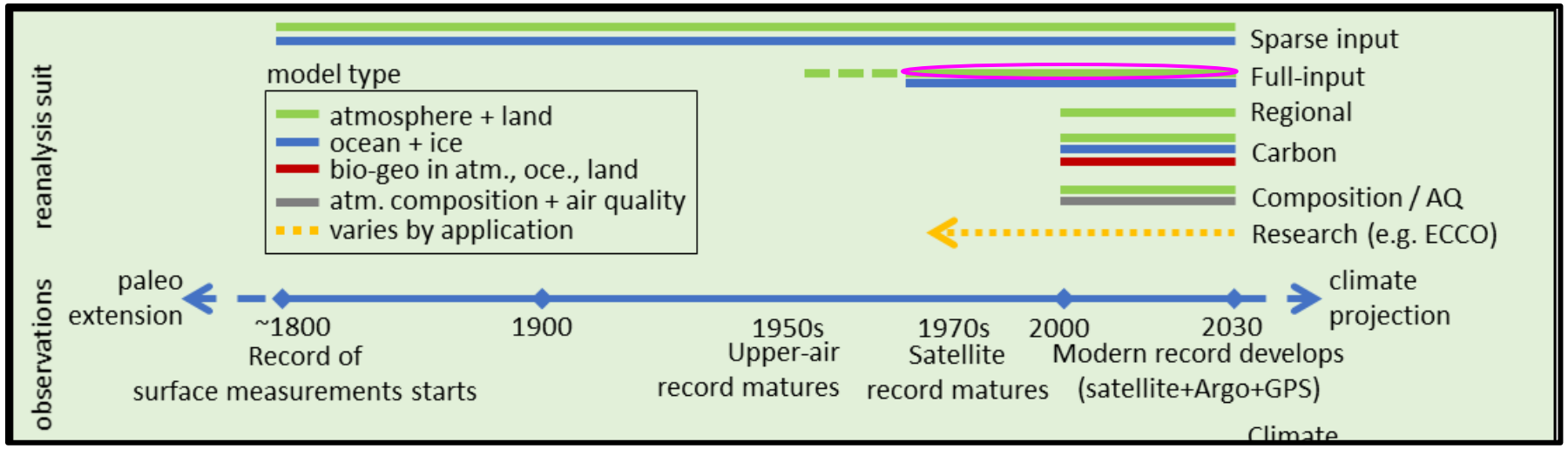


- Homogenized data.
- High scientific quality of data.
- Constrained by observations.
- Ease of access (ARCO).

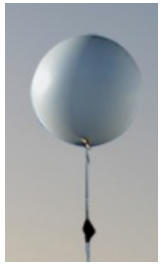
Well-maintained reanalyses (and ensemble reforecasts) are well-suited for training of DL models

ERA5 Schematic

Recommendation for the future U.S. suite of reanalysis

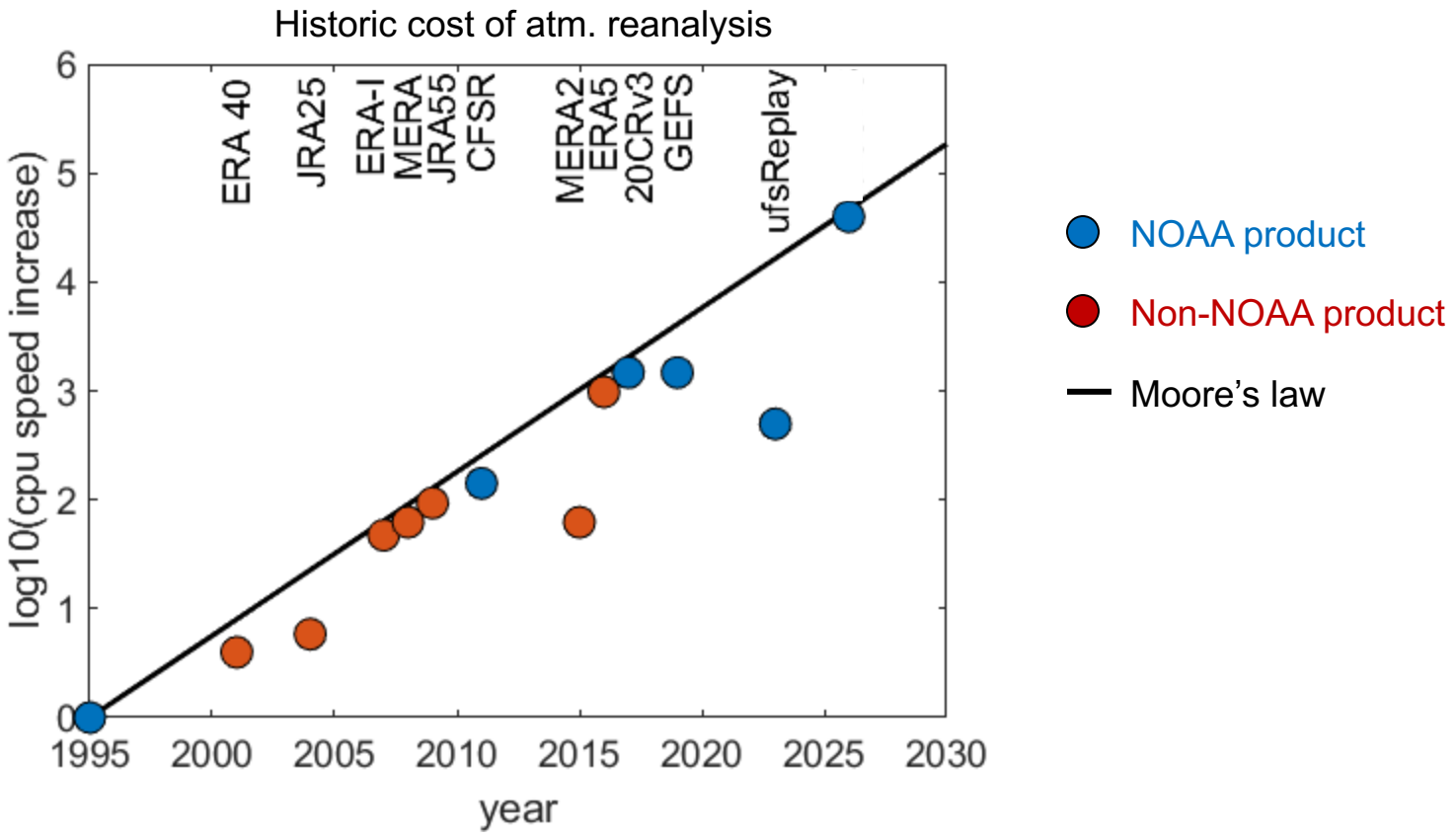


ERA5



- Reanalysis is limited by historic data availability.
- Recent workshop recommended a suite of reanalysis to maximize data impact.
- ERA5 is only a small part of the full suite of products.

Brief history of reanalysis datasets



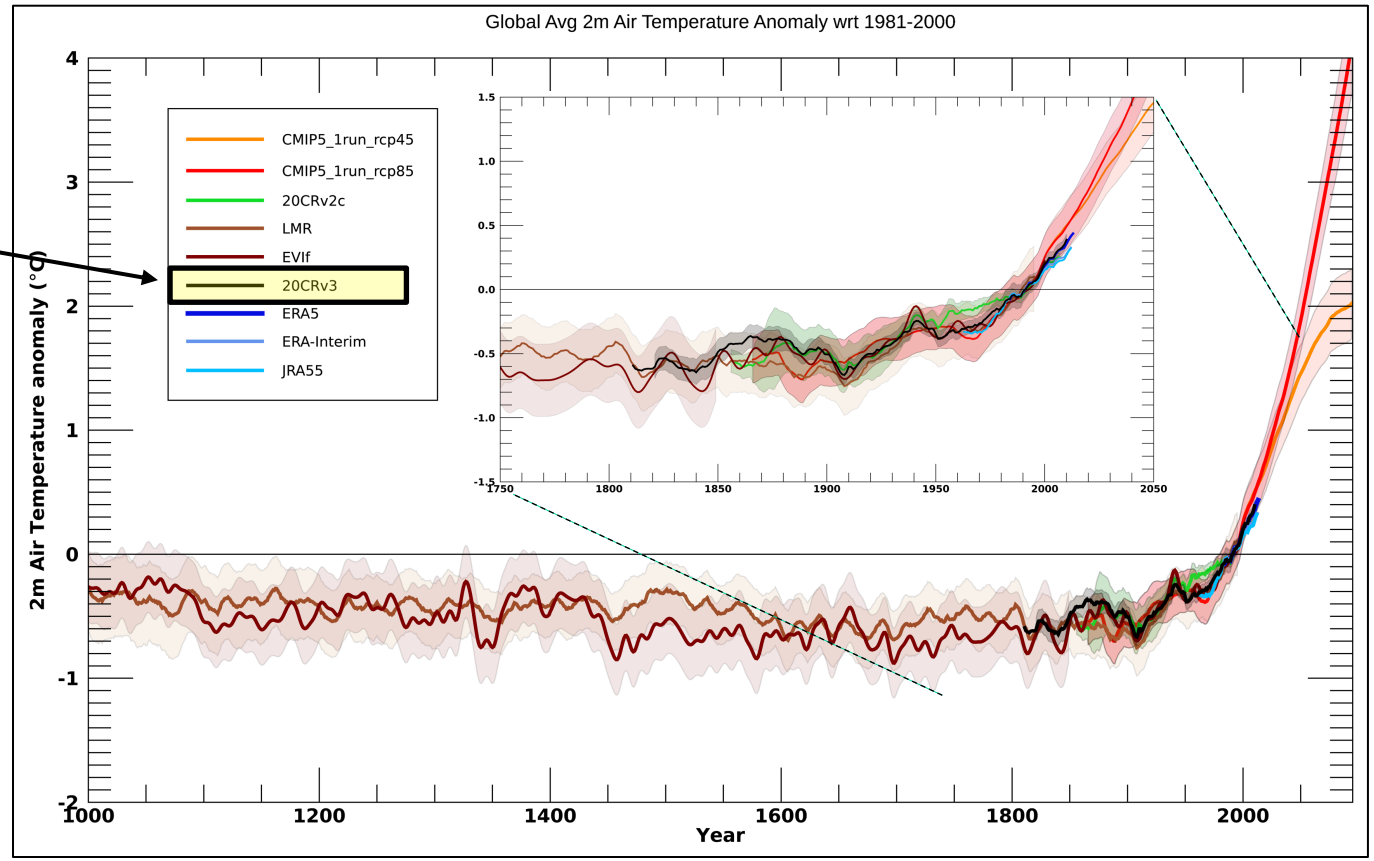
- Over the last 3 decades, NOAA produced multiple reanalysis products.
- The latest three products (20CRv3, GEFSv12, UFS-replay) are recommended as ML training datasets.
- However, it is essential for NOAA to invest into a modern native reanalysis (e.g. UFS-R1)

Brief history of NOAA reanalyses datasets

- NCEP/NCAR-R1 (1995): One of the original reanalysis.
  - Still running in production 30 years later!!!
  - Not recommended for use
- CFSRv2 (2010): First coupled reanalysis
  - Still in production and operational
  - Not recommended for ML training
- 20CRV3 (2017): 200 years of reanalysis with 80 members
  - Best centennial reconstruction available.
  - By design not as accurate as ERA5 during modern period.
- GEFSv12 (2019): Designed for reforecast experiments
  - Best used for ensemble reforecasts (training and evaluation).
- UFSReplay (2023): "replay" of the coupled UFS model to ECMWF reanalyses
  - **Recommended NOAA dataset for atmosphere, ocean, ice, and land ML training.**

20CRv3: 200 years of reanalysis with surface pressure observations

20CR is completely independent from *any* land surface temperature observations



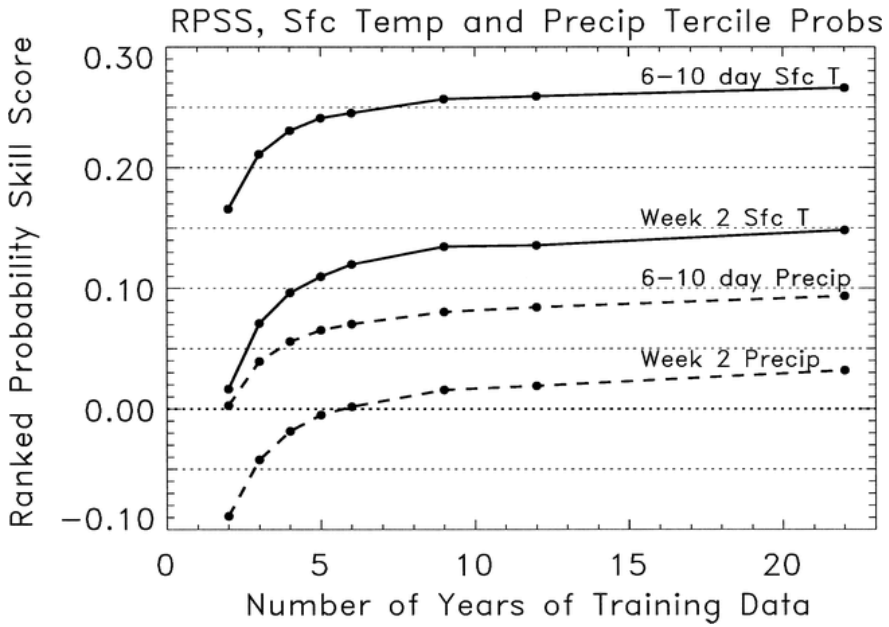
Reanalyses provide an instrument-based link between paleo reconstructions and climate model projections

Continuous reconstruction of weather:

- 80 ensemble members at 75km resolution
- Assimilated observations are limited to surface pressure measurements.
- 200 years of data.
- Available as file downloads.

Figure courtesy of Compo et.al

GEFSv12 reforecast

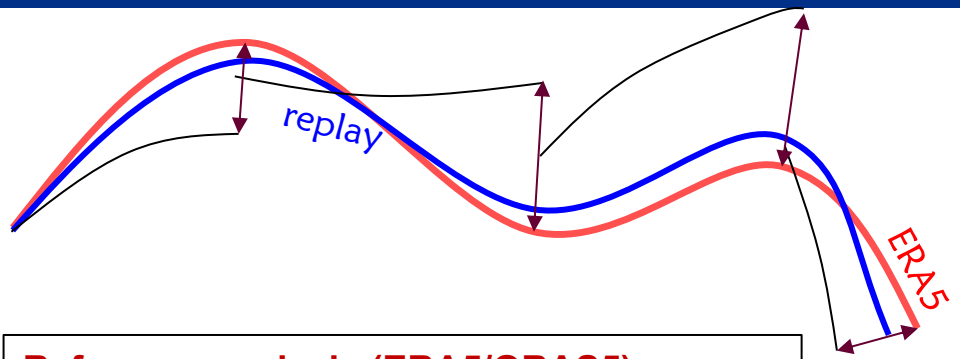


Hamill 2004: At least 10-20 years of reforecast data is needed to effectively post-process forecasts of near surface quantities.

- Target audience: developing post-processing tools for real-time forecasts (configured similar to operational GEFSv12 but at a lower resolution).
- Fixed period of time (2000-2019) 80 members at 75 km resolution.
- 31 ensemble member forecast to 16 days 4x day, to 35 days 1x day.
- Available on AWS

Figure from Hamill 2004

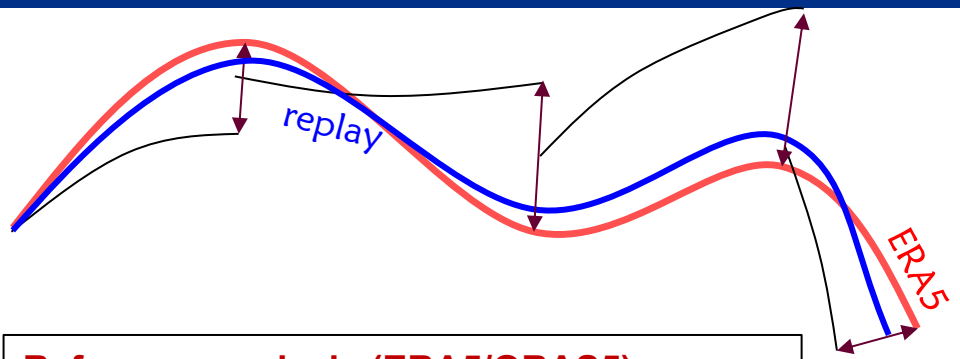
UFS Replay



Reference analysis (ERA5/ORAS5)
First forecast
Increment from ERA5
Replayed forecast forced by the increment

- Latest NOAA UFS model is replayed (nudged) to existing reanalysis (ERA5/ORAS5).

UFS Replay



Reference analysis (ERA5/ORAS5)
First forecast
Increment from ERA5
Replayed forecast forced by the increment

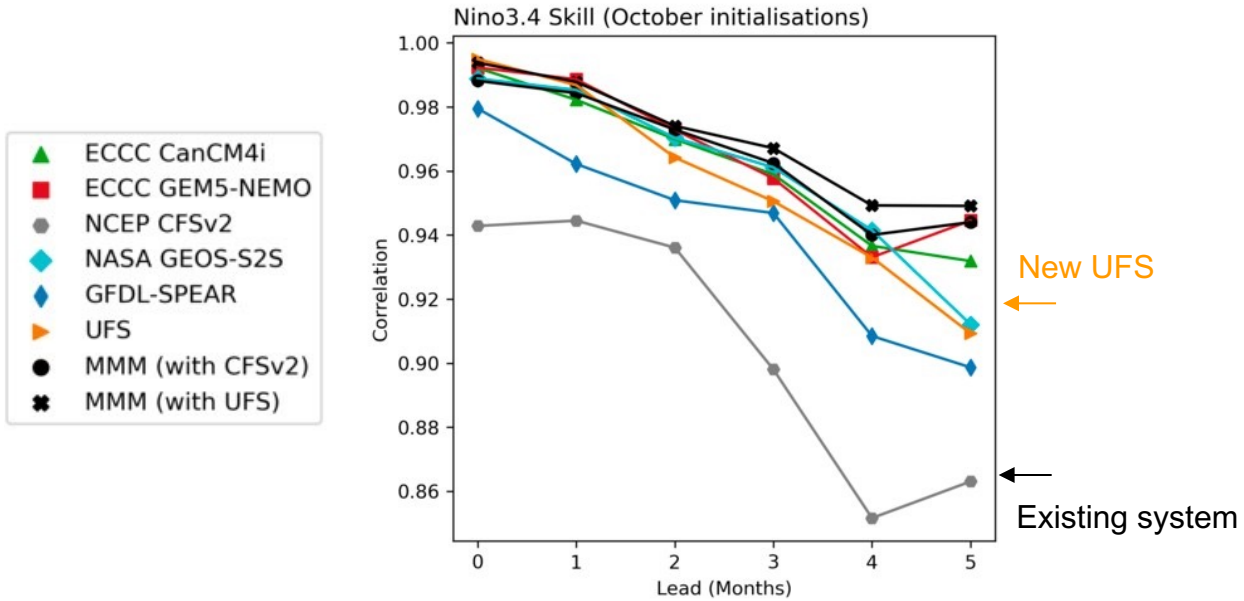


Figure courtesy Newman

- Latest NOAA UFS model is replayed (nudged) to existing reanalysis (ERA5/ORAS5).
- Latest model:
  - Atmosphere (FV3), latest atmospheric physics (HR1), MOM6, CICE6, NOAH-MP land, WWIII waves.
- Period:
  - 1994-present.
  - Output 3h on native model levels.
- Quality:
  - Indication of significant improvement over existing NOAA models.

ARCO storage

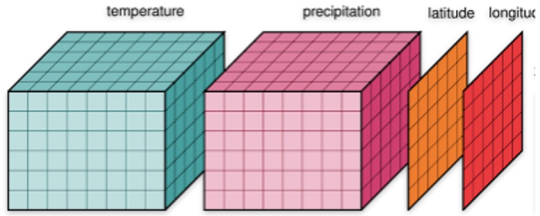
(1) Model-native

archive of netcdf/grib2/bufr files (~1M files)

(2) ARCO

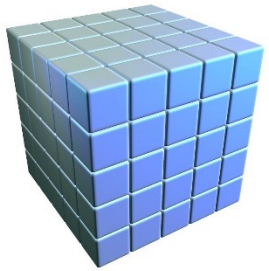
ZARR archive:

- logically one file
- stored in small chunks that are easy to access

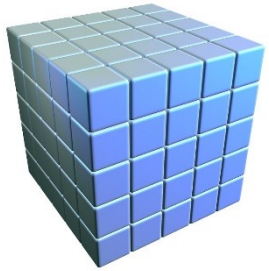


(3) "GPU"-ready

Training input



Training targets



- Traditional files in traditional storage:
  - Great for on-disk access.
  - Terrible for internet access (e.g. ftp/S3).
  - Often can be behind a firewall or on tape.
- ARCO– analysis ready cloud optimized
  - E.g. zarr on the cloud.
  - Very efficient access to subsets of data through cloud-optimized chunking.
  - Originally a Python-native package.
- Training-ready dataset:
  - Specific to the configuration of the ML experiment.
  - Optimized for ingest into GPU/TPUs.
  - Typically has to be optimized by individual training teams.

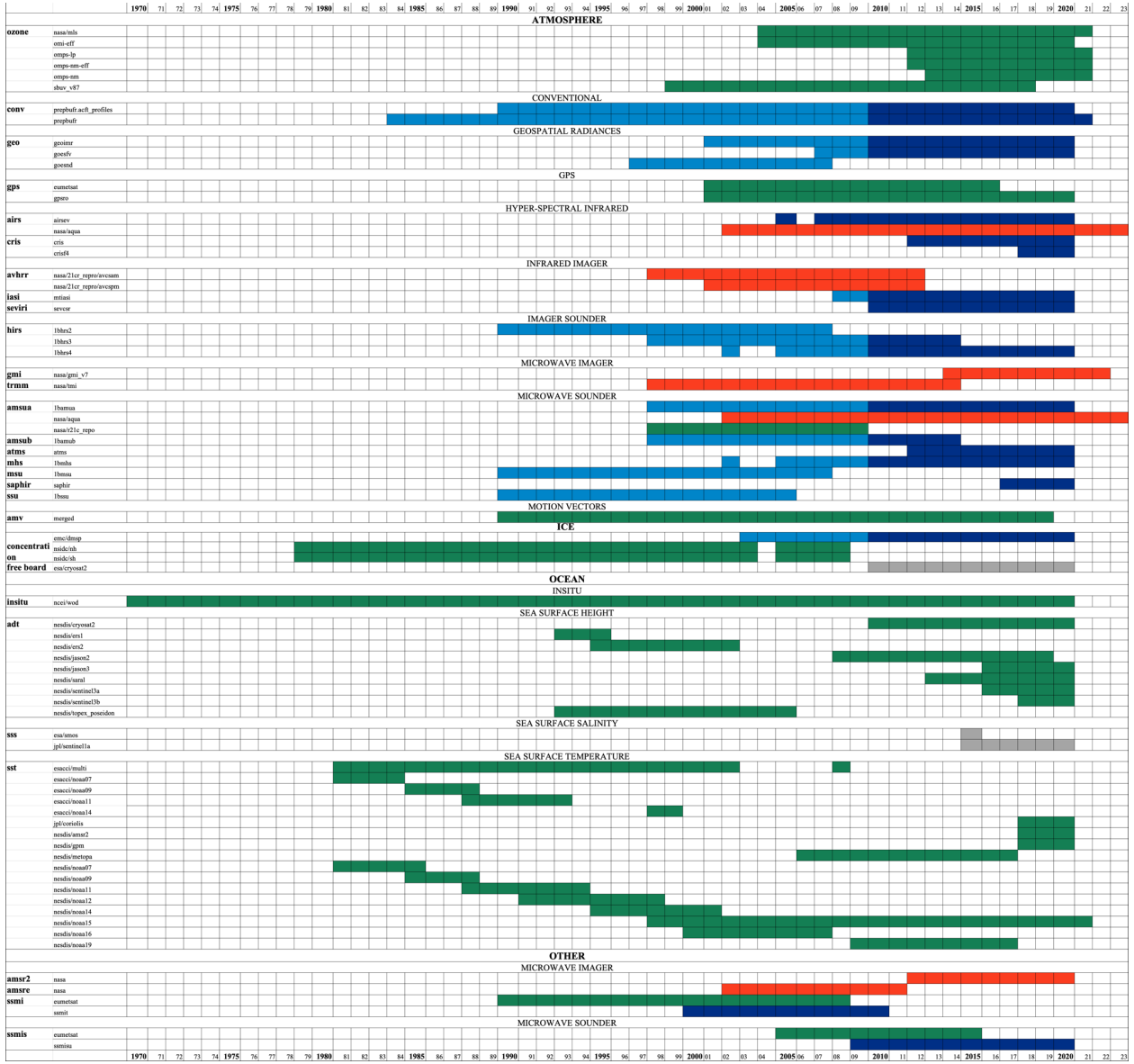
Availability of UFS replay data

- Native model output:
  - AWS S3 (about 1 Pb).
  - https://noaa-ufs-gefsv13replay-pds.s3.amazonaws.com/index.html
- ARCO representation for training of model emulators:
  - Key dynamic and microphysics variables needed for training a model like GraphCast.
  - Atmosphere on gaussian grid and native model levels.
  - Ocean and ice on native (tri-polar) model grid and levels.
  - https://console.cloud.google.com/storage/browser/noaa-ufs-gefsv13replay
- ARCO representation for data science:
  - To be reformatted in Q1 2024 and stored on Azure.
  - Commonly used 2D variables (2mT, precipitation, 10m wind, ….).
  - Focused on coupled processes.

Other US-centric dataset

- Catalogue of NOAA data on AWS:
  - https://registry.opendata.aws/collab/noaa/
- AMIP runs from NOAA GFDL.
- GFS/GEFS operational analysis and ensemble forecasts
- Convective allowing datasets over CONUS
  - HRRR (3km) real-time products (back to 2014 with inconsistent versions).
  - CONUS404 (NCAR/USGS): 40 years at 4 km downscaling of ERA5.
- NOAA-NASA observational archive for reanalysis (1979-near real time).

NOAA-NASA observational dataset



- Observational inputs to the reanalysis.
- Covers a full range of observations for: atmosphere, ocean, ice, and land.
- Restricted data removed.
- Observations stored in non-ARCO files.
- Need a lot of curation and documentation to be useful to non-specialists.
- Still been developed.

Possible future datasets

- Replay with nested model:
  - NOAA is developing a nested configuration of UFS that will allow replay generation with local refinement over CONUS (e.g. 13 or 3 km depending on configuration)
  - Some production is possible in the next 1-2 years

- UFS R1 – native reanalysis with the NOAA UFS model and data assimilation
  - In development (production is possible in 2-3 years from now).
  - No computational resource is identified yet.
  - Effort under resourced (about ½ of the Copernicus investment into ERA6).
  - Prime opportunity for AI enhancements.

Concluding remarks

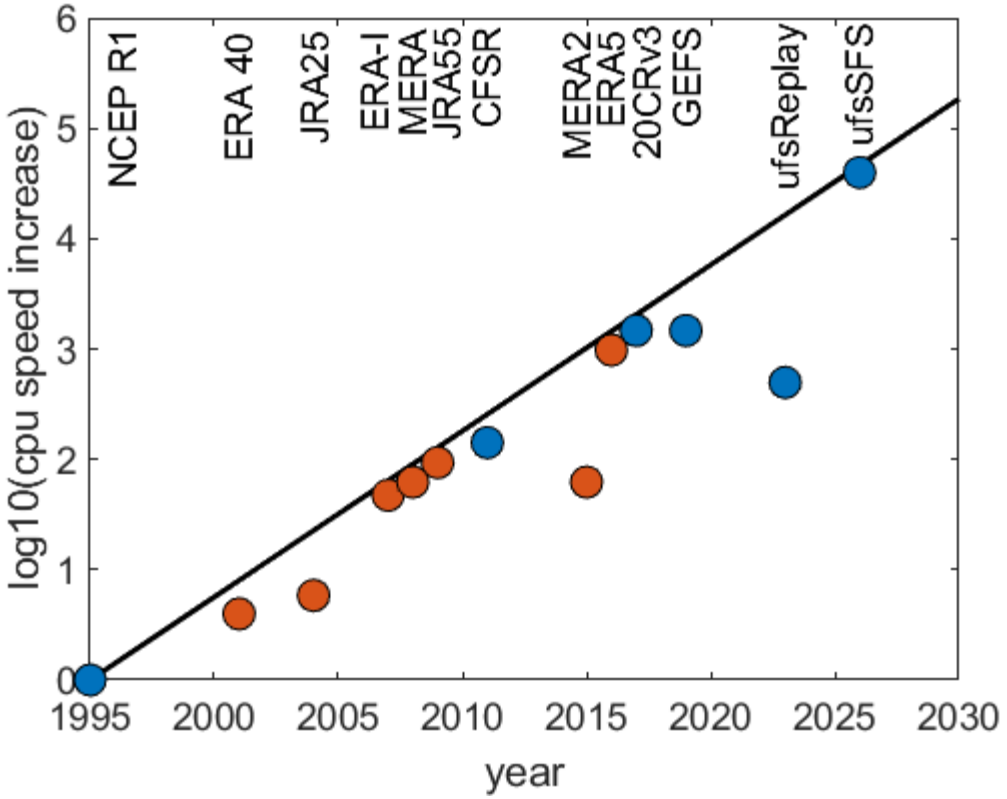
- NOAA has a wide variety of dataset for ML training:
  - Less organized compared to the Copernicus datasets (e.g. ERA5).
  - More modern access than Copernicus.
  - Varied degrees of quality.

- Opportunities for improvements:
  - Develop open ML ecosystem around existing NOAA data.
  - NOAA has to develop in-house expertise in ML model development and training using NOAA data to stay relevant.

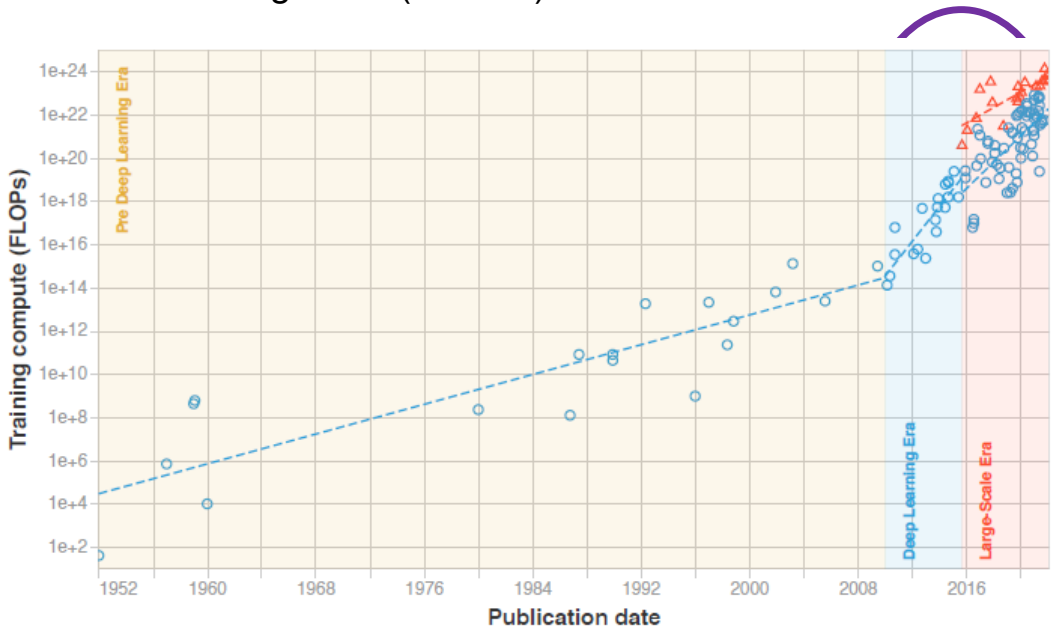
- Essential need for NOAA-native reanalysis:
  - Should be produced with ML learning in mind.
  - Key link between NOAA science and any future implementation of AI emulators in operations.

Role of ML in the future reanalysis and NWP

Historic cost of atm. reanalysis



Training costs (FLOPS) of milestone ML models



Doubling time

pre-deep learning 21.3 months

Deep learning 5.7 months

Large scale models 9.9 months

- To date, reanalysis development was constraint by Moore's law.
- **Combination of hardware, software, and science is accelerating ML development** significantly faster than the Moore's law.
- Can reanalysis and data assimilation for NWP benefit from the acceleration in ML science?

https://arxiv.org/pdf/2202.05924.pdf