

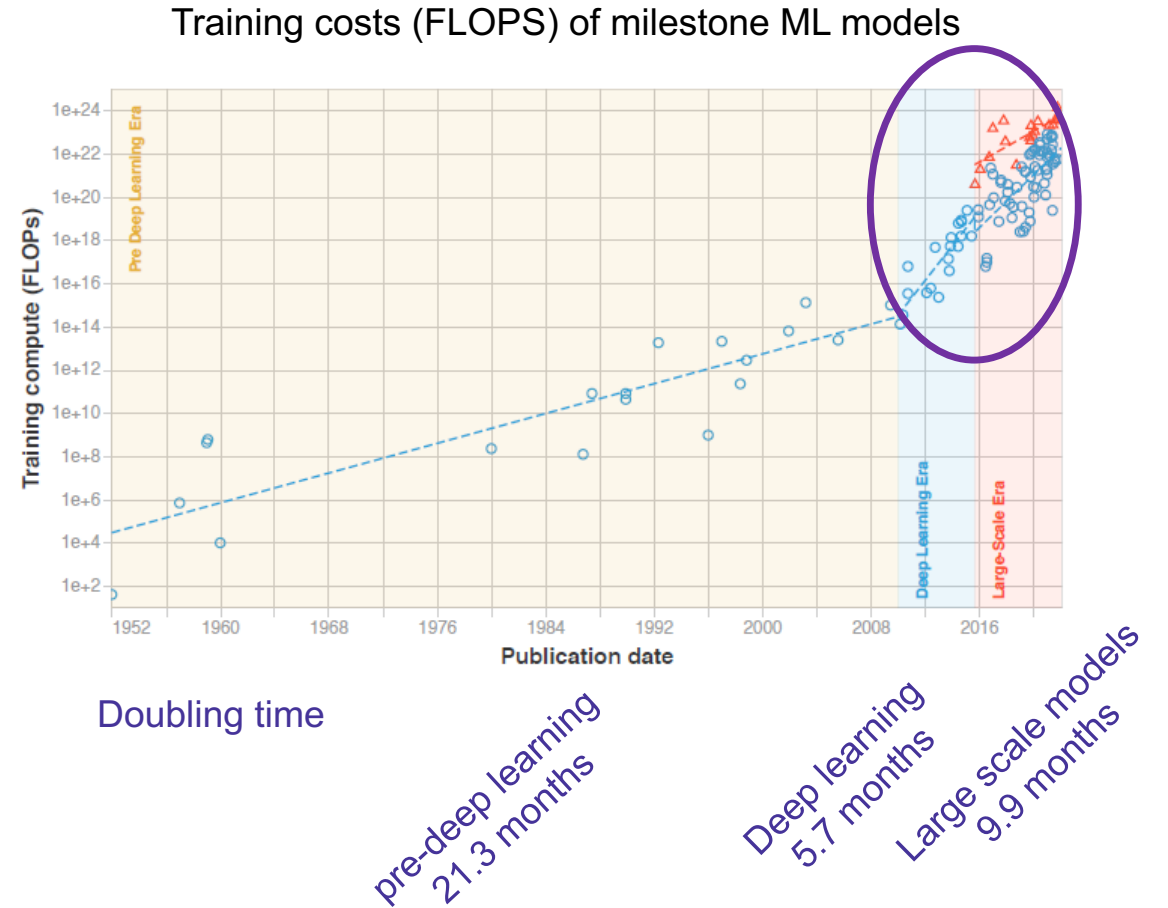
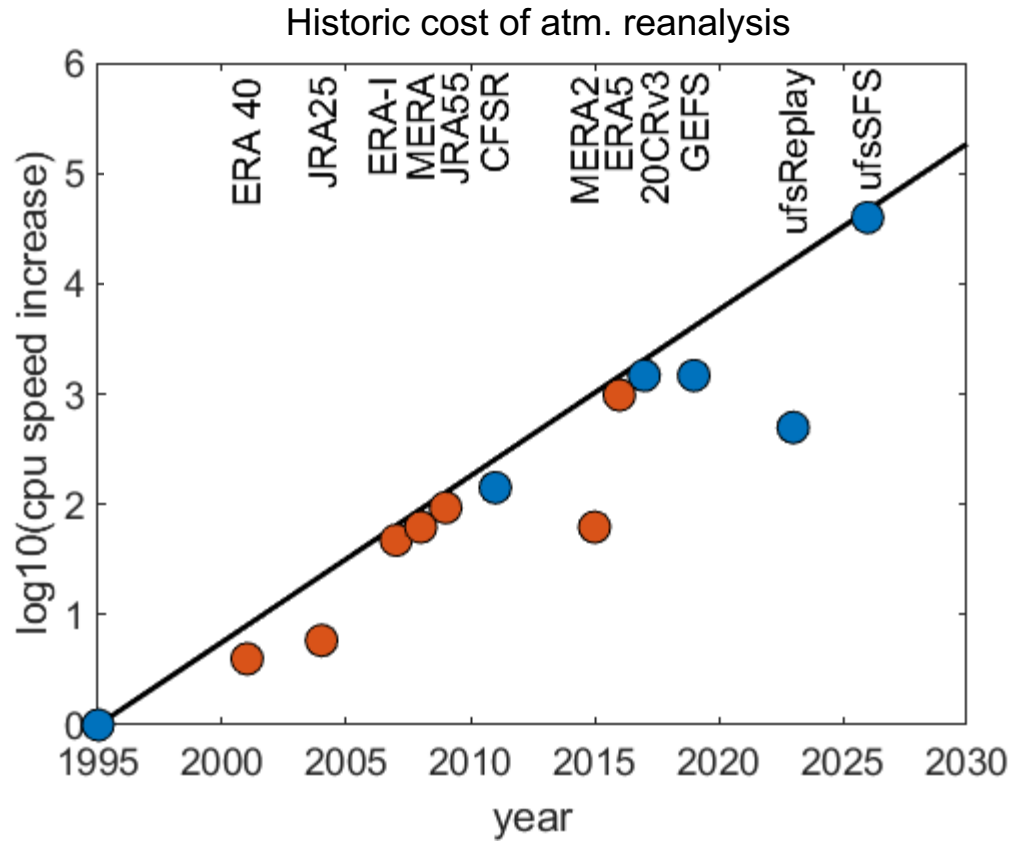


Research direction: reanalysis

Sergey Frolov

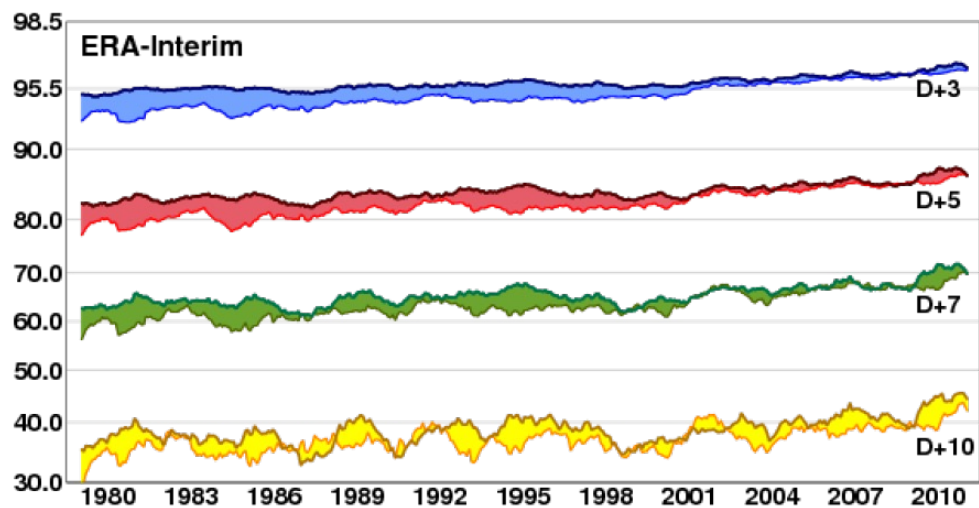
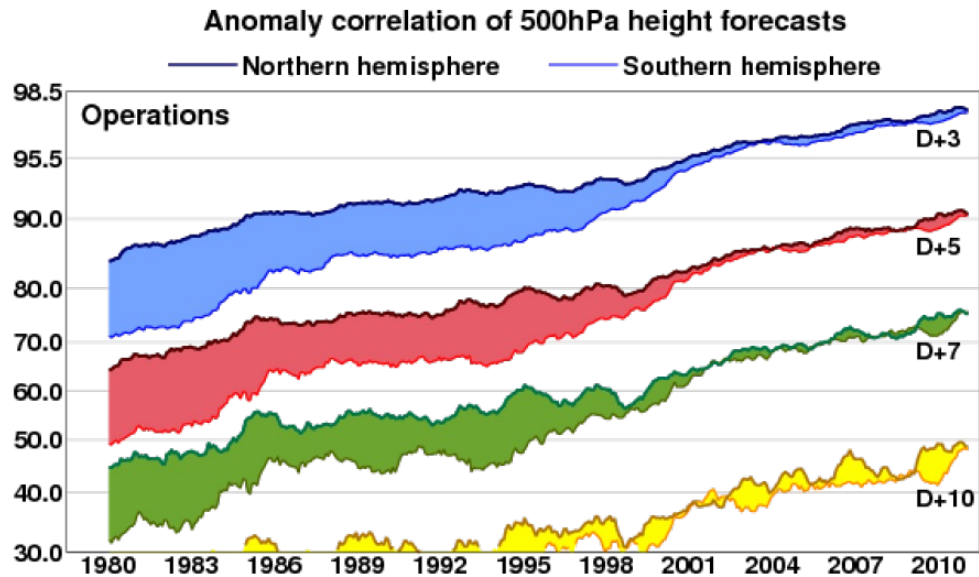
Presented at: AI4NWP workshop
November 29, 2023

Role of ML in the future reanalysis and NWP



- 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?

“Normal” reanalysis workflow

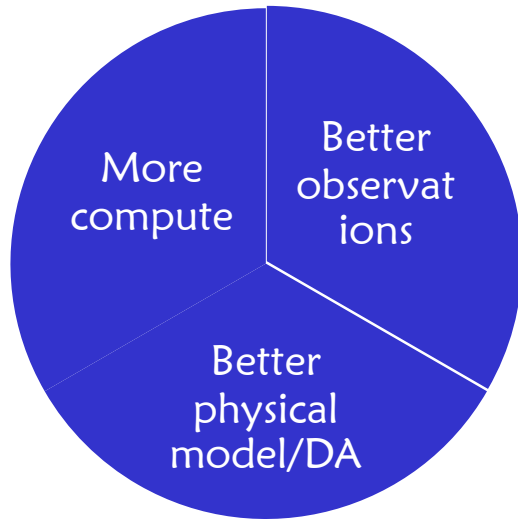


- A complete reanalysis is produced every 7-15 years using a frozen version of the model/observations/DA system.
- Reanalysis is a great improvement over historic archive of operational analysis.
- Cost of reanalysis production is staggering (~\$10-\$100M)
- The wall time for generation is long: takes years to produce.

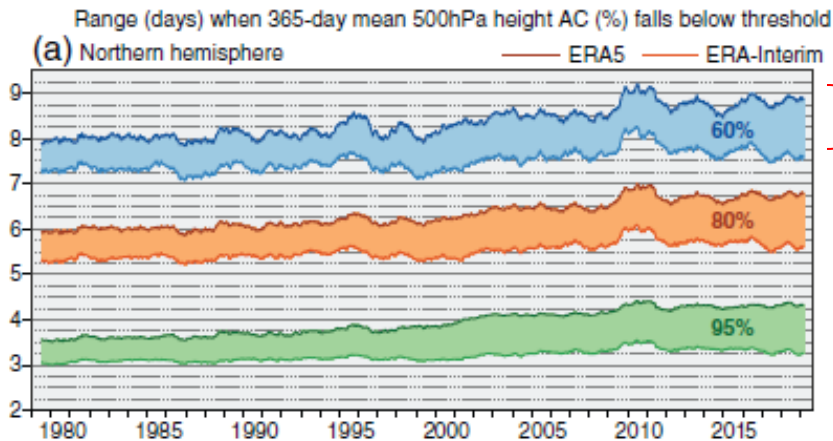
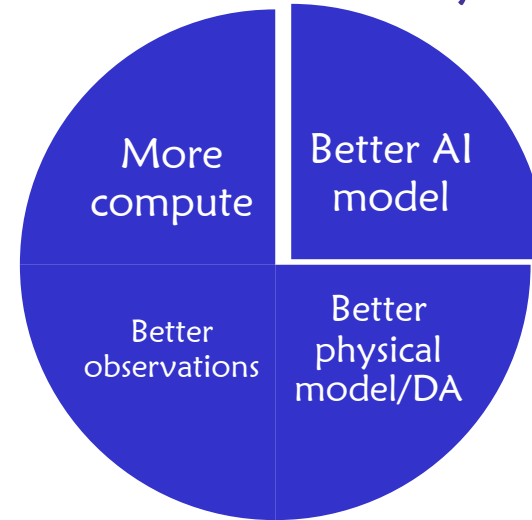
- It is likely that ML training will become the main “user” of future reanalysis
 - Could it also serve as a pathway for transitioning model improvements to operational forecasts.
- Under this new concept of operations, existing practice of reanalysis production becomes unsustainable.

More reasons to change the way we generate the reanalysis

Current reasons for new versions of reanalysis



Future reasons for new versions of reanalysis



Skill gain from 10 years of progress in model, DA, observations and compute

If AI models become an integral part of the reanalysis production, then improvements in AI models alone might trigger re-production of the reanalysis.

Circular reasoning fallacy

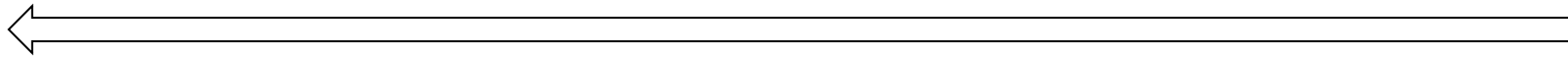
Instead of offering evidence, the argument simply repeats the conclusion, rendering it logically incoherent.



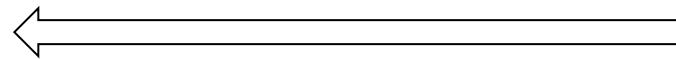
- Using AI for production of reanalysis is separate from using reanalysis for training of the AI.
- How can we do both?

Hierarchical reanalysis

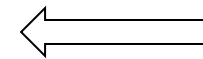
Long re-analysis to pre-train models
(50 year, every 6 hours, 25km)



Medium length focused on quality
(20 years, every 1 hour, 9 km)



Short focused on model updates
(2 years, every 15 minutes, 1 km)



- What is the needed length of the datasets above?
- Can we tune ML models trained at coarse resolution and cadence on a shorter dataset with higher resolution and cadence?
- Can we iterate between steps (1)-(3)?

Please discuss options for small/medium/large projects

- What are minimum viable research efforts to advance the science of reanalysis and their use for ML training?
- Effort size:
 - Small: ~1 person years (e.g. evaluation of forecasts from a single emulator)
 - Medium: ~3 person years
 - Large: ~10+ person years (e.g. original GraphCast development)