The interplay between data assimilation and artificial intelligence

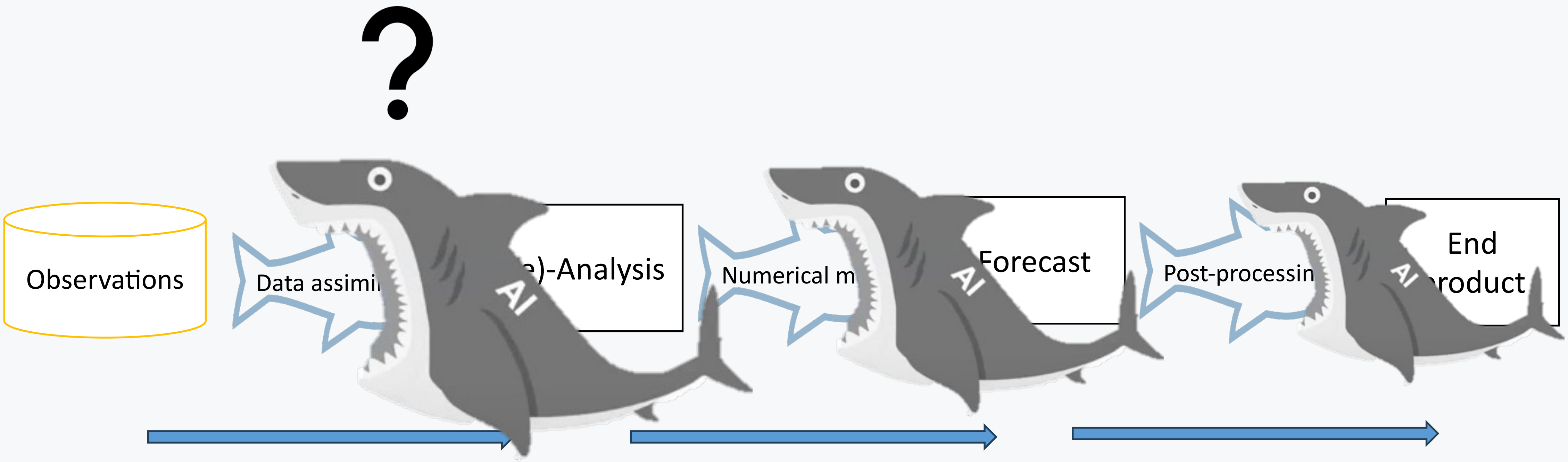
Julien Brajard– Julien.brajard@nersc.no

NERSC
STIFTELSEN NANSEN SENTER
FOR MILJØ OG FJERNMÅLING
NANSENSENTERET · BERGEN · NORGE

Contributors: **Yue (Michael) Ying**, Laurent Bertino, Leo Edel, Anton Korosov, Zikang He, Fabio Mangini,

Integrated Forecasting System

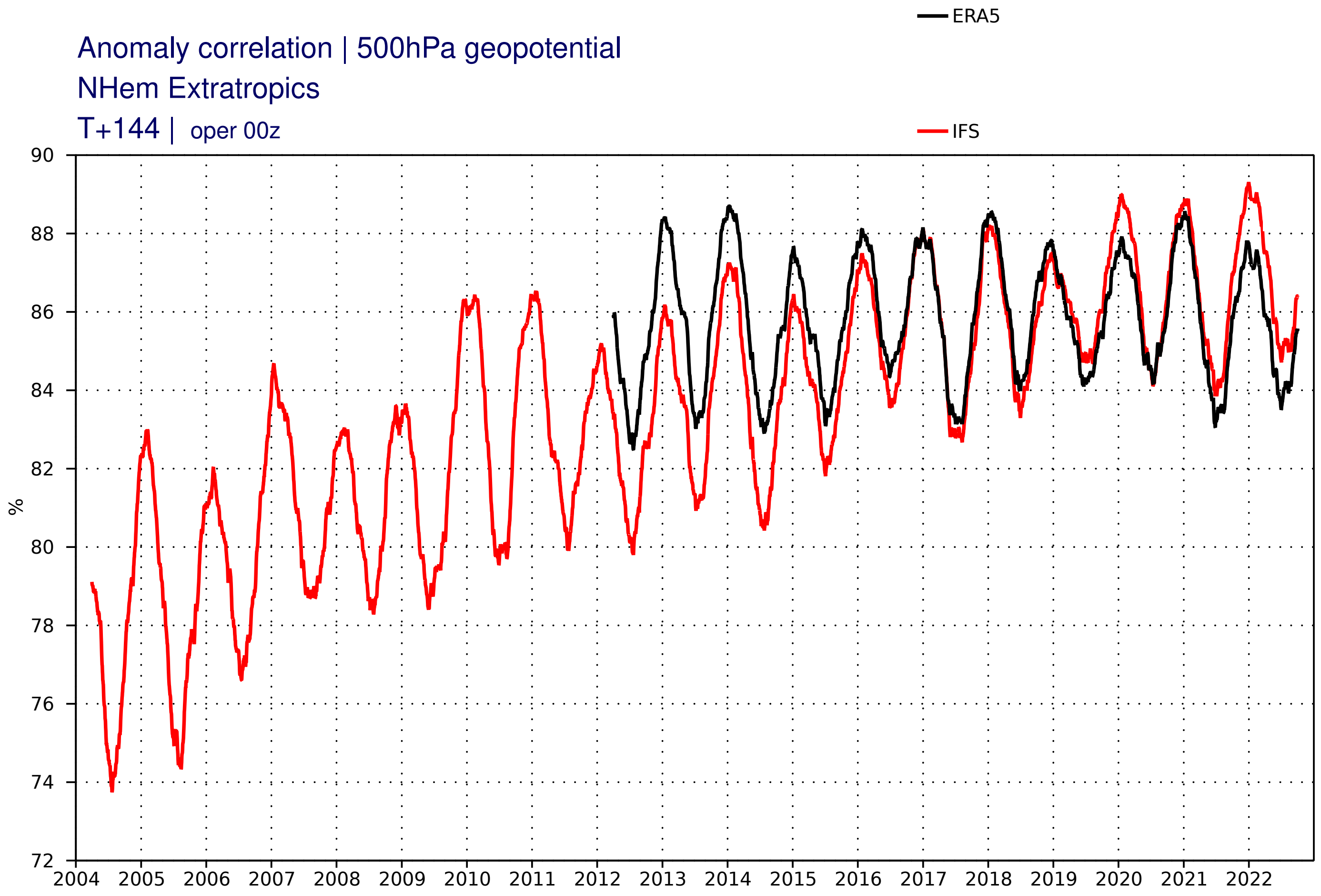
NERSC



Courtesy of Stephan Rasp

ISDA online webinar, Jan. 2024

Achievements of Machine Learning in Numerical weather forecast



AIFS: data-driven model

Lang et al. 2024

Principle of a data-driven model

NERSC

Physical-based model:

Physical principles / equations:

$$\frac{d}{dt}\mathbf{v} = -2\boldsymbol{\Omega} \times \mathbf{v} - \frac{1}{\rho}\nabla_3 p + \mathbf{g} + \mathbf{F}$$

Discretization/Numerical schemes

$x(t)$ → Physics-based model → $x(t + dt)$

Data-driven model:

Physics-based model

observations

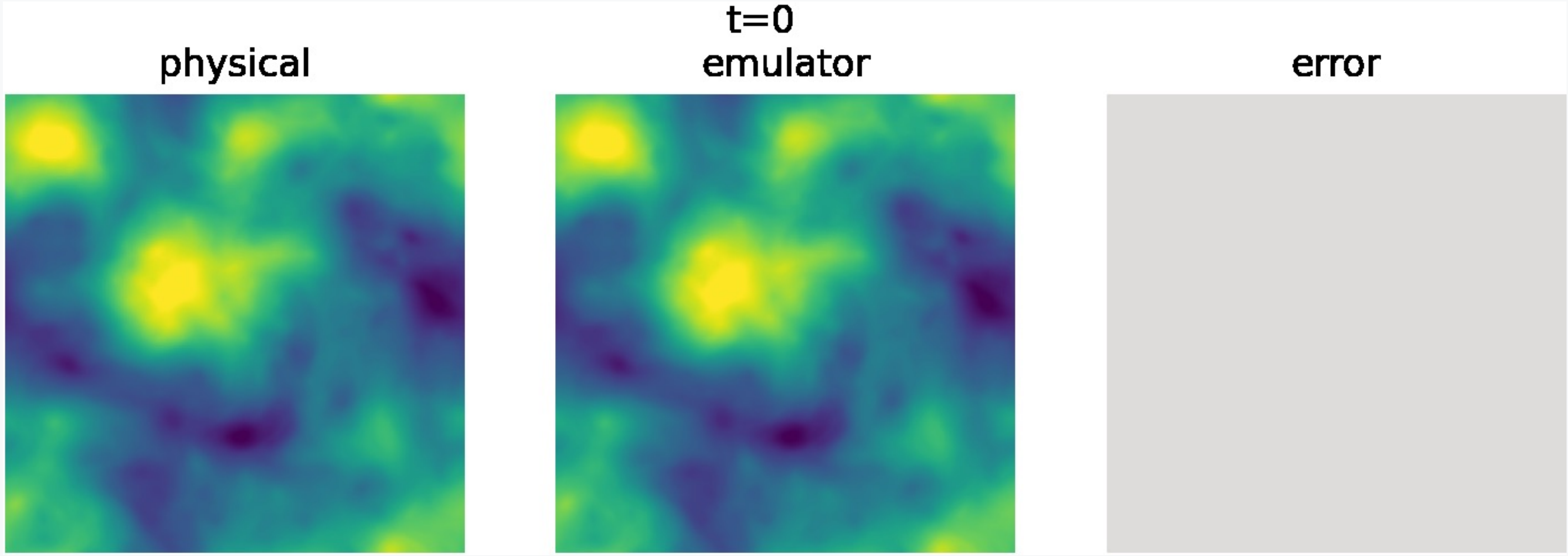
Data

Training

$x(t)$ → Data-driven model → $x(t + dt)$

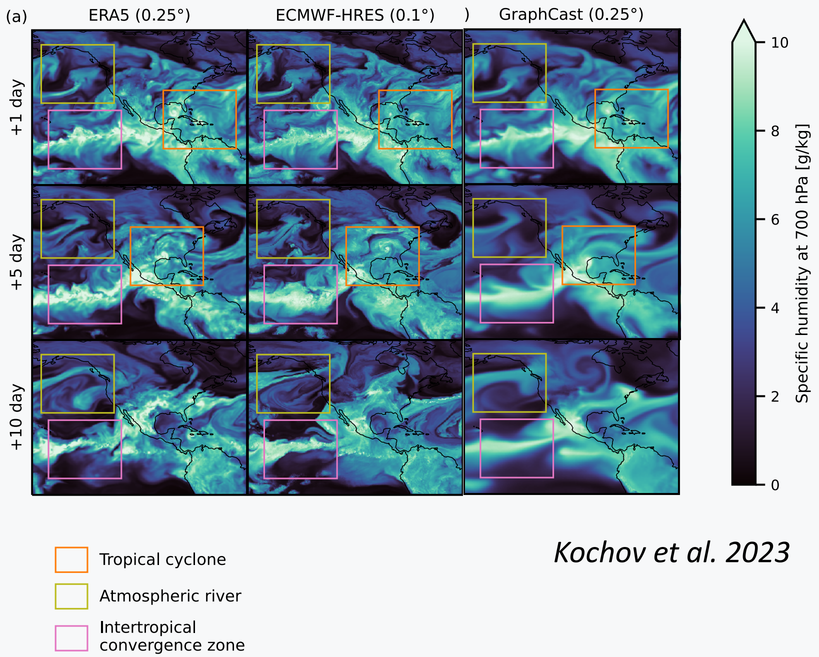
Example of a full data-driven model (emulator)

Simple emulator of a quasi-geostrophic model (low-complexity system)



Training is simple to setup, but…
Blurring effect
Also (not shown): instability on long run

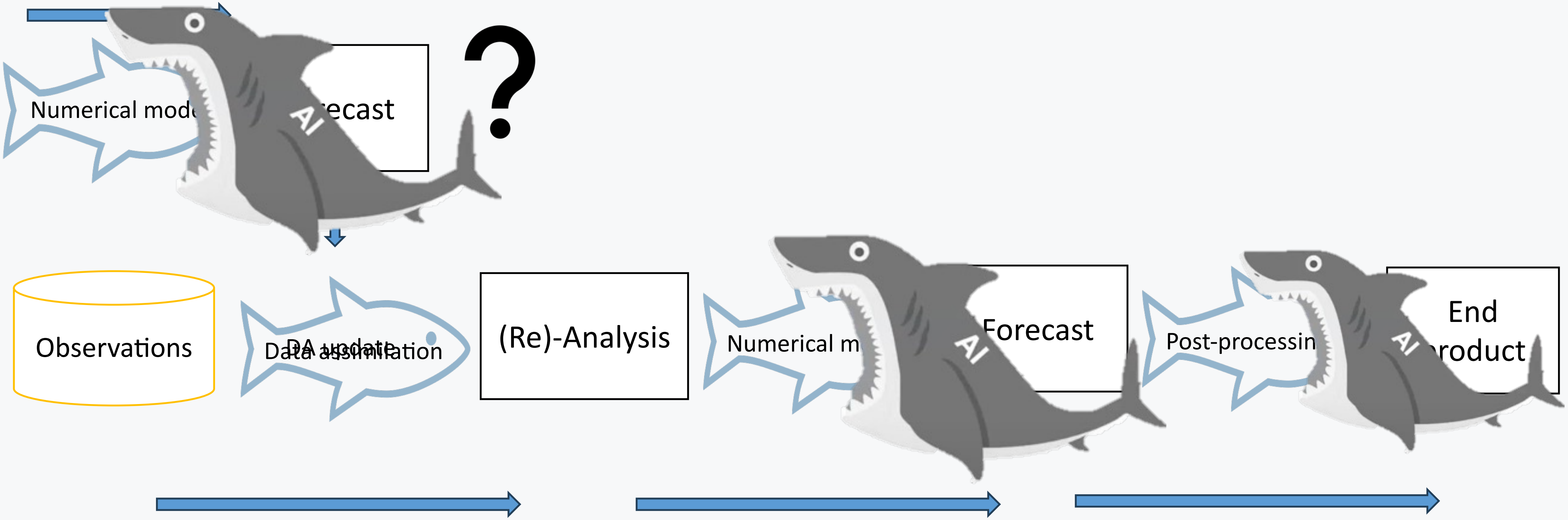
In collaboration with Yue (Michael) Ying



Kochov et al. 2023

HRES: hours of computation with hundreds of CPU
GraphCast: < 1 min on a single Google TPU

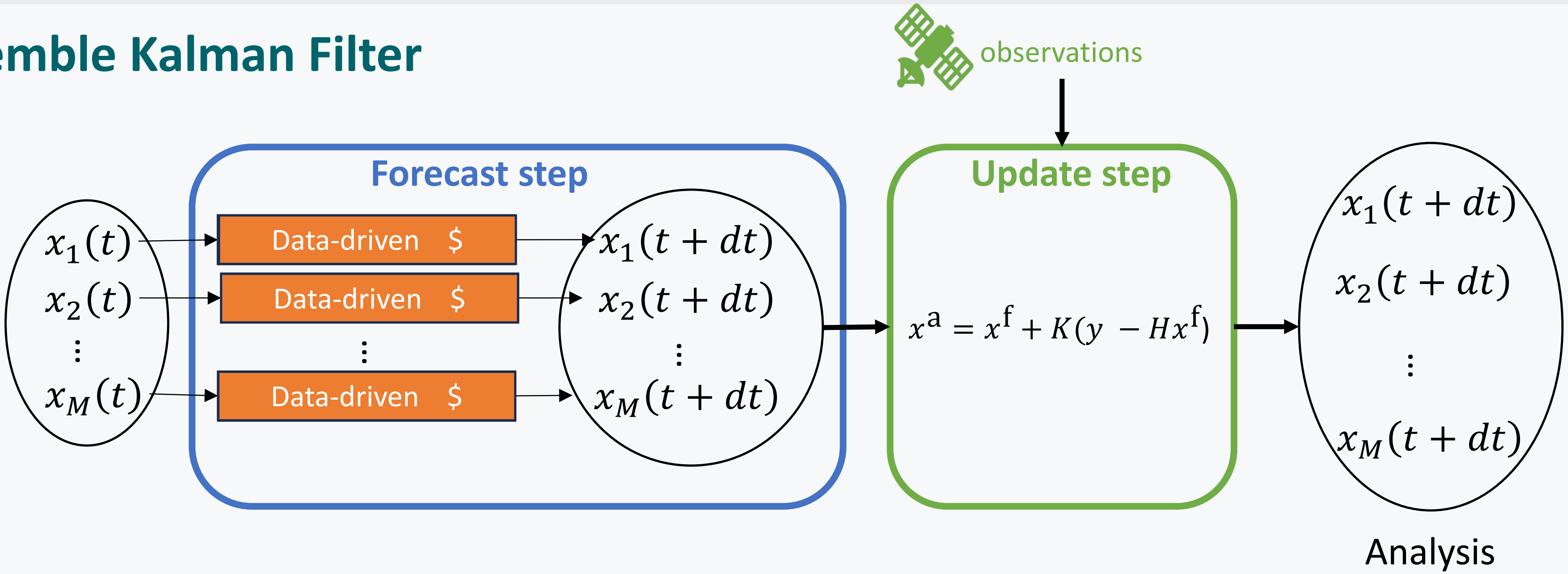
What does it have to do with data assimilation?



Courtesy of Stephan Rasp

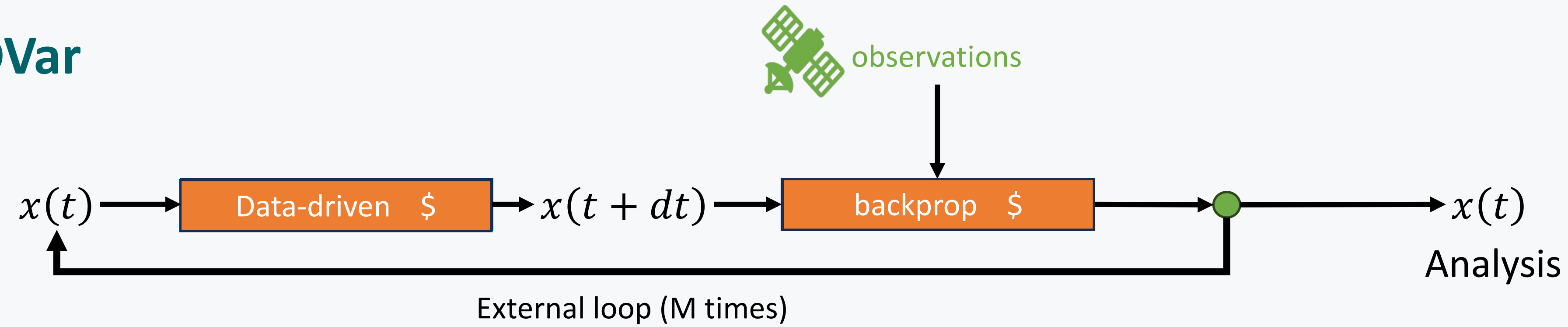
Benefits for EnKF and 4dVar

Ensemble Kalman Filter



See Eliott Lumet's talk

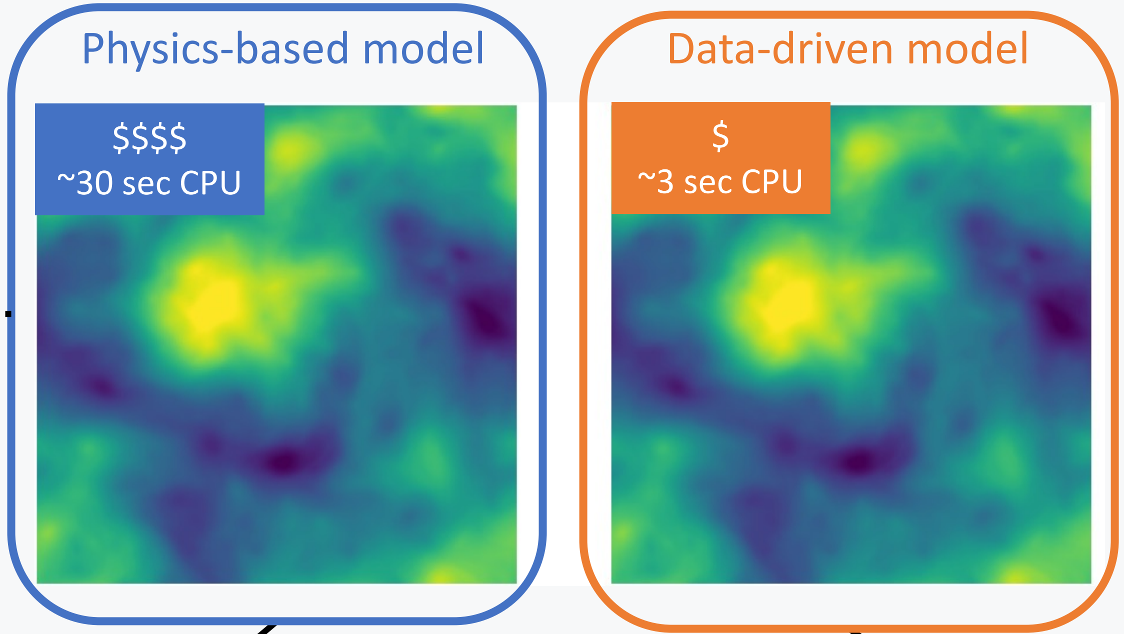
4DVar



FengWu 4DVar, Xiao et al., 2023
Durand et al. 2024,
https://zenodo.org/records/11636139

Illustration

Twin experiment setup: physics-based model is used to produce observations



synthetic observations

EnKF

EnKF

Physics-based analysis

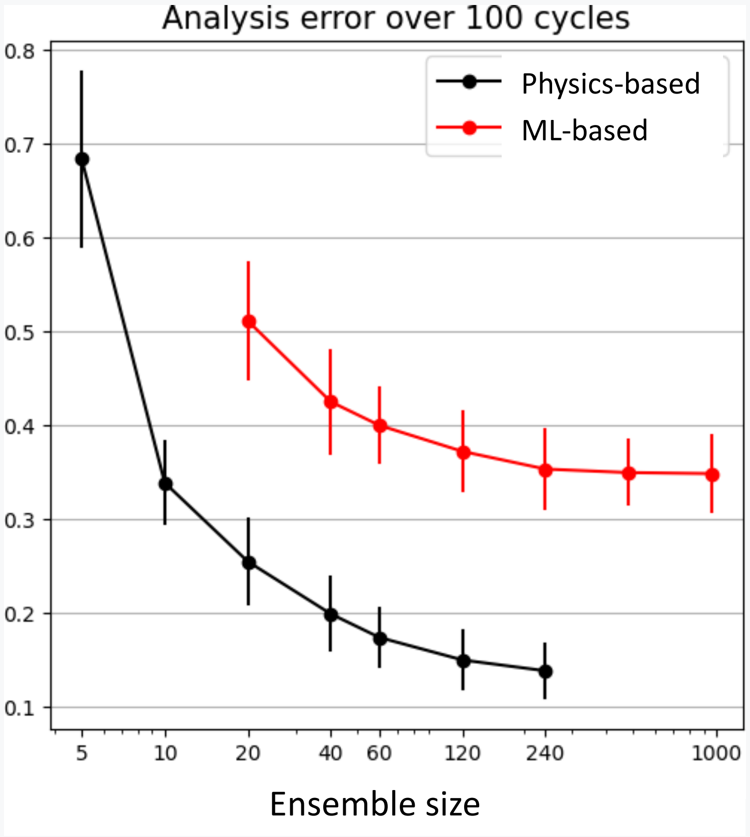
ML-based analysis

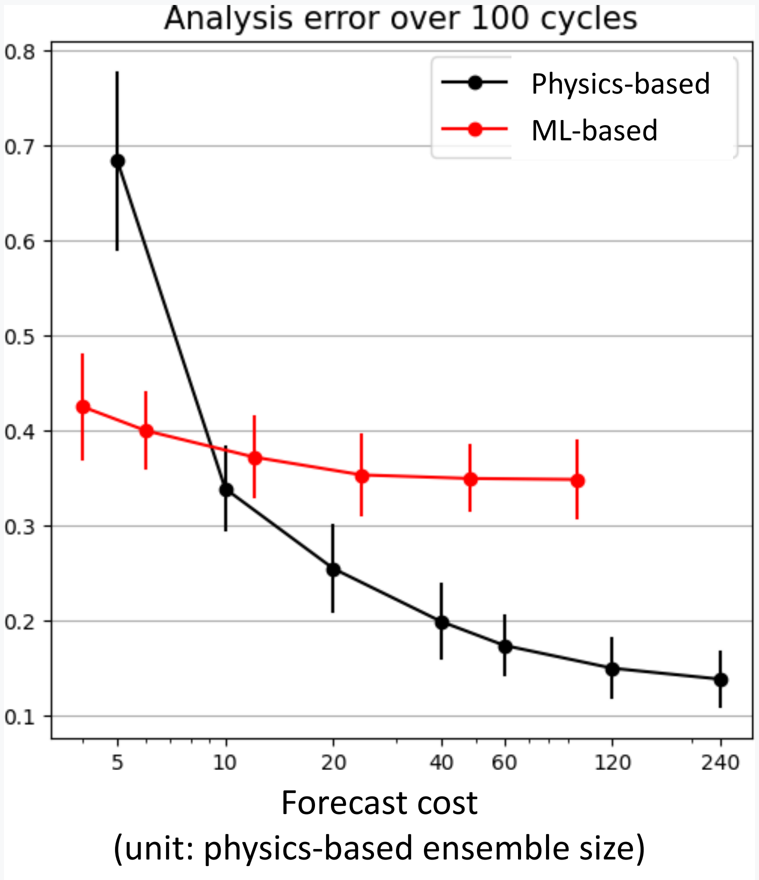
Experiment implemented using
Keras and NEDAS (~2 weeks)

https://github.com/nansencenter/NEDAS/tree/develop

Results

NERSC





DA with emulators can be beneficial in case of limited computational resources

Comment:

- ✓ Physics-based model is perfect (no model error)
- ✓ Due to blurring, more inflation is needed in the ML-base experiment to maintain the spread

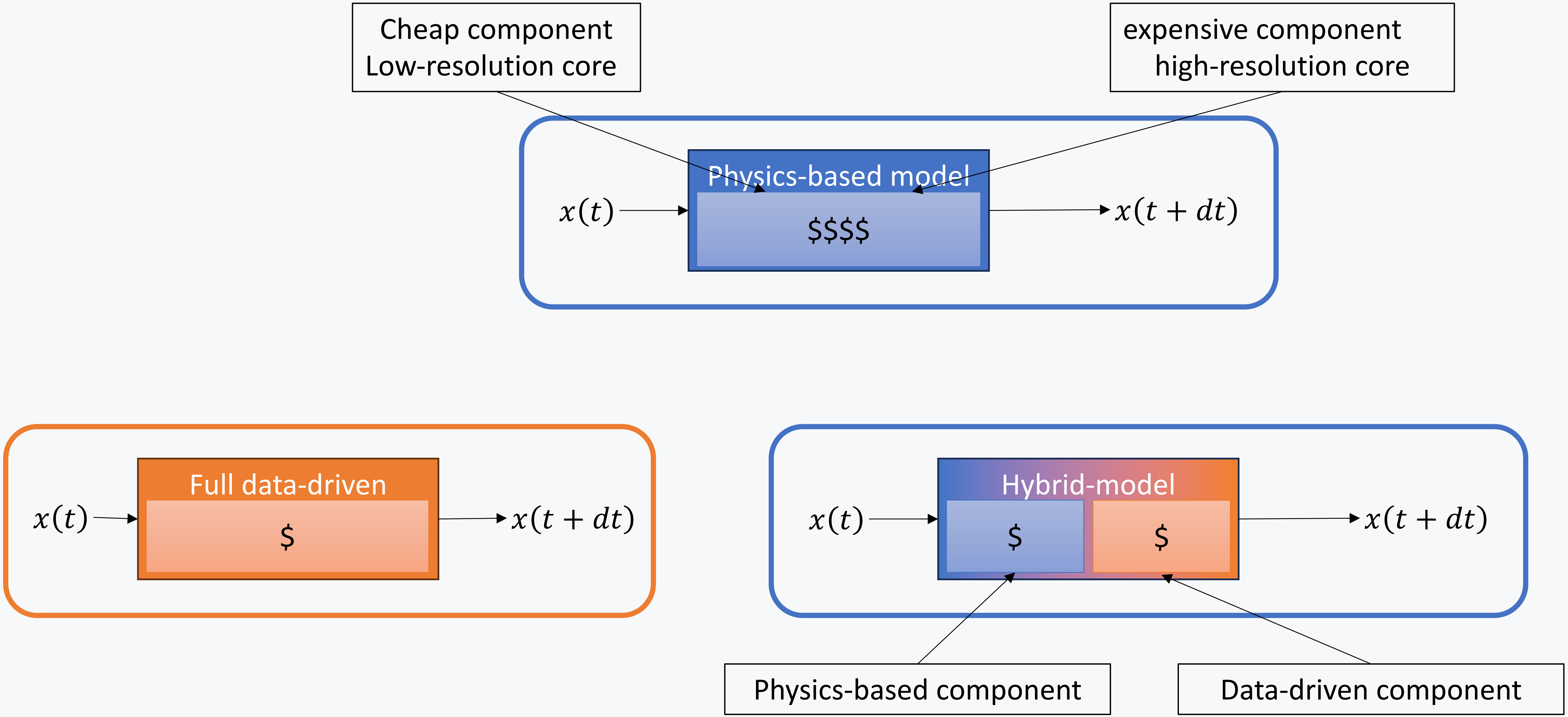
Y. Ying

How to overcome this underestimation of spread (i.e. blurring)?

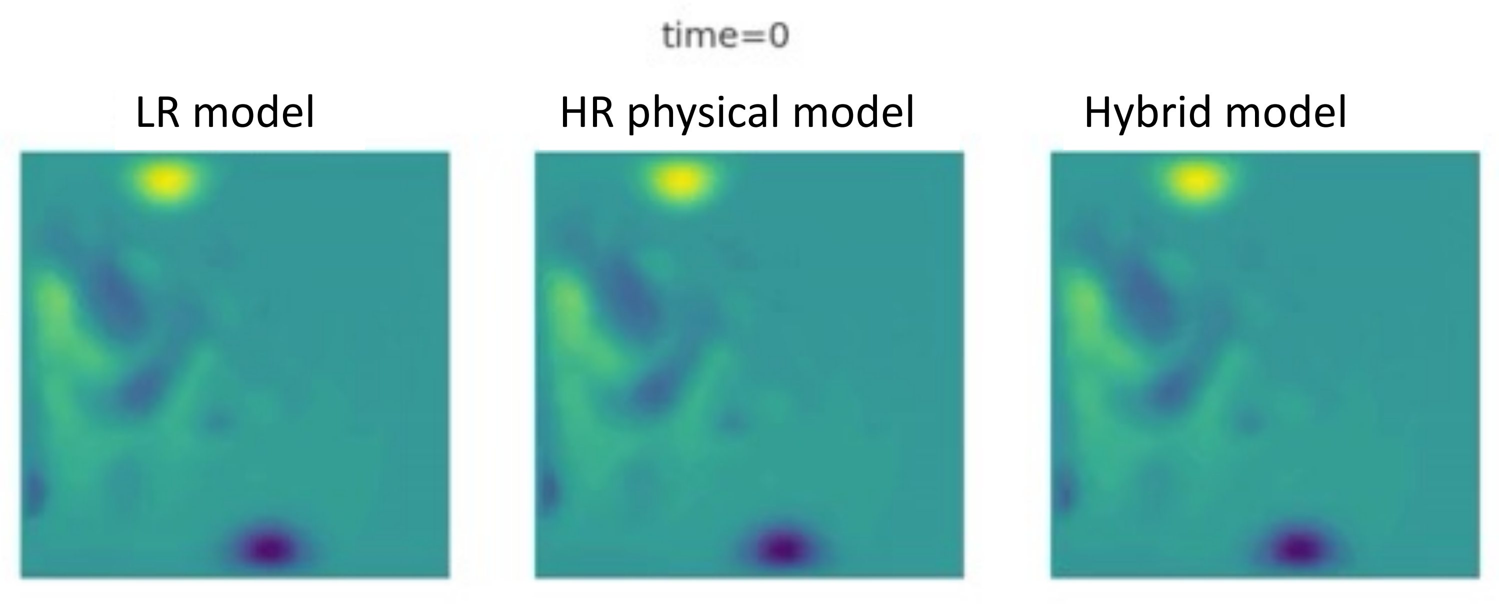
Idea #1
Hybrid models

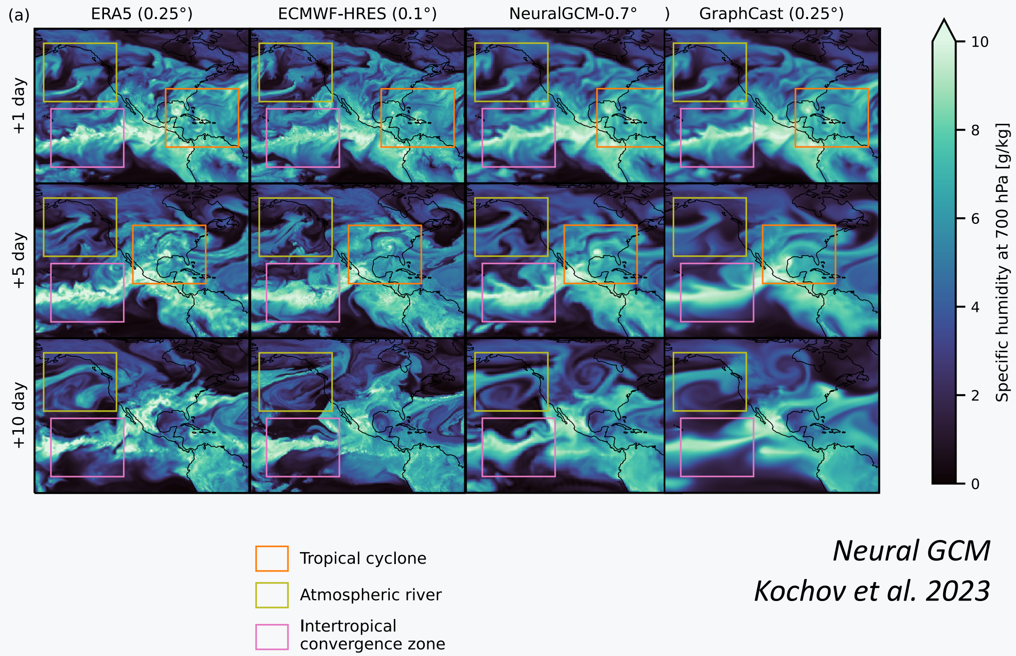
Idea #2
Generative models

Dynamical models and Machine learning



Example of an hybrid model



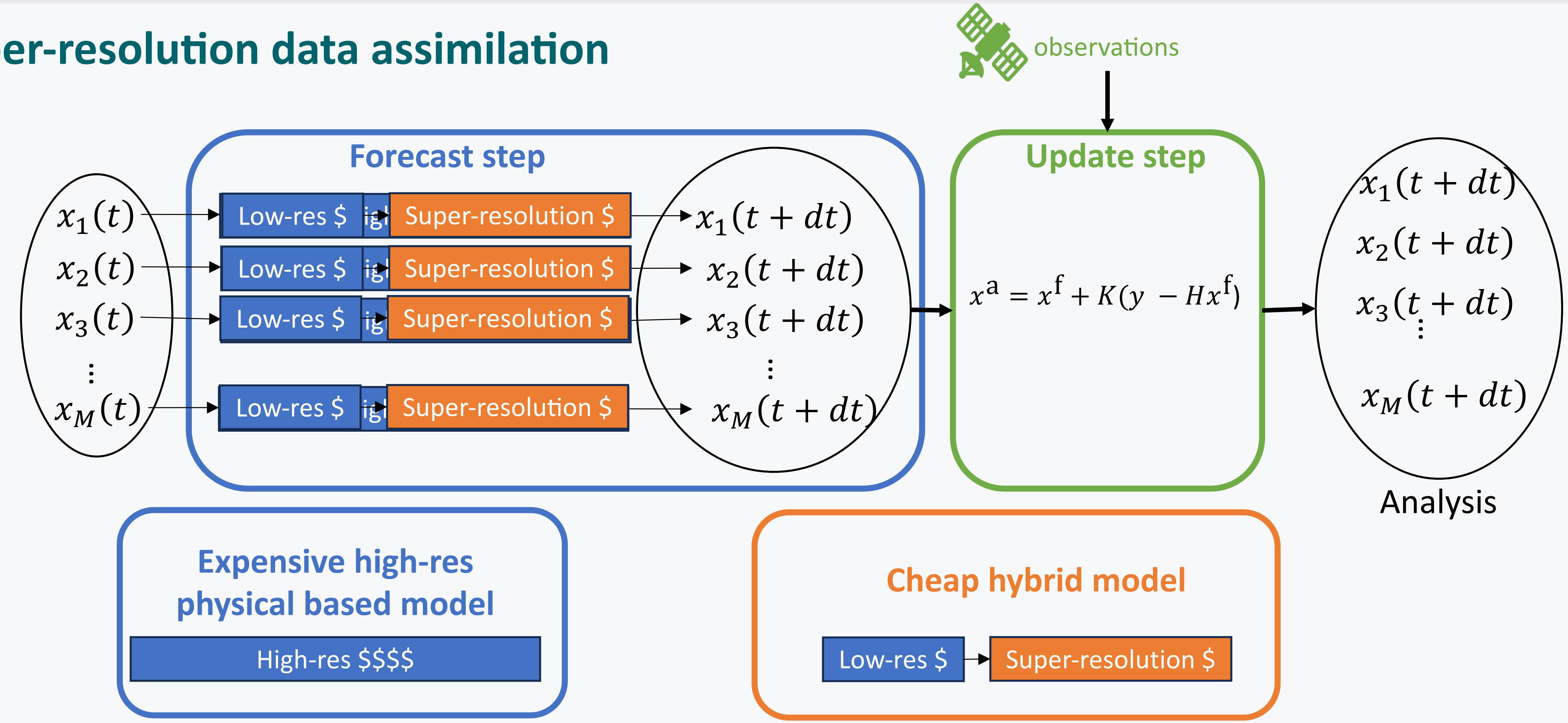


Neural GCM
Kochov et al. 2023

No smoothing
Stable on long lead time

SuperResolution data assimilation

Super-resolution data assimilation



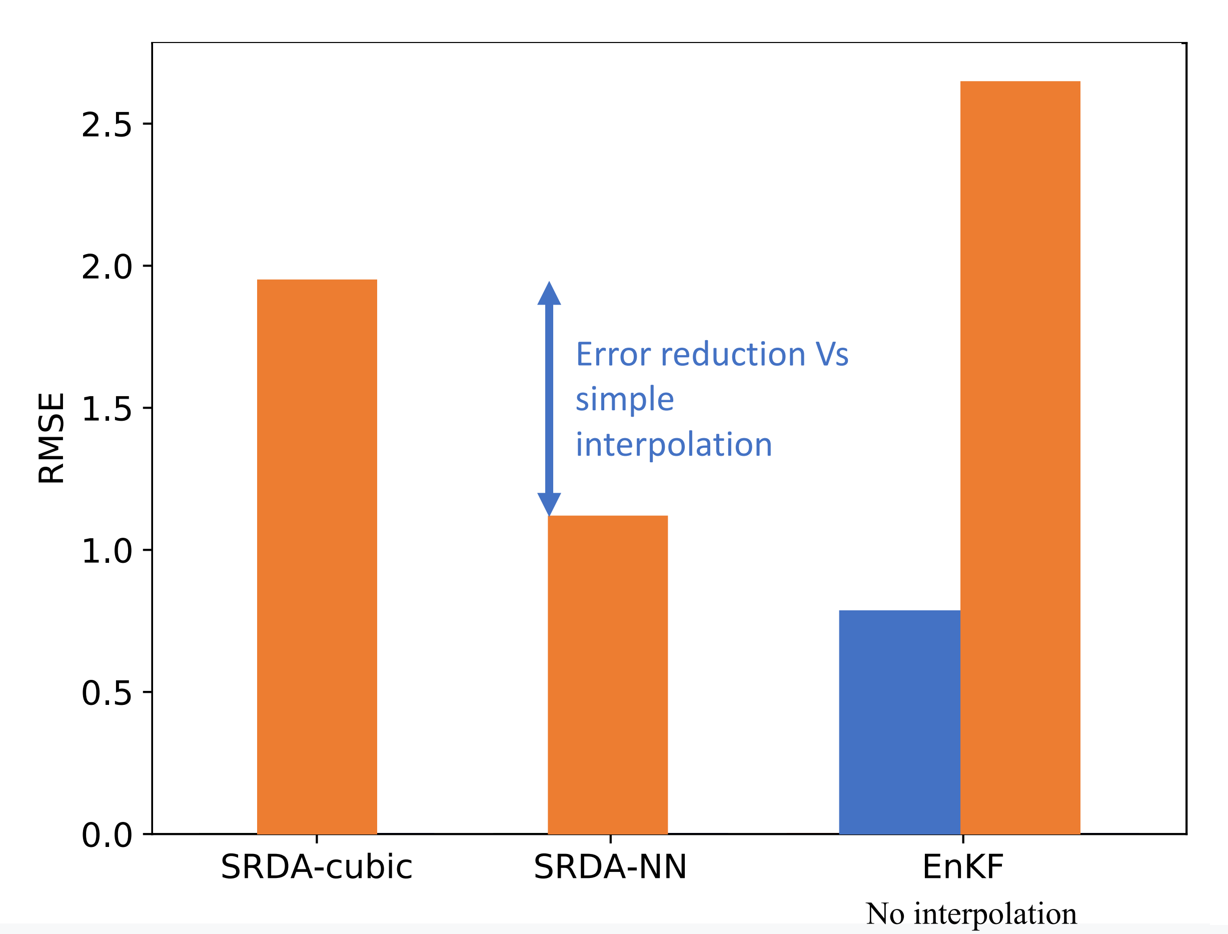
Barthélémy et al., 2022

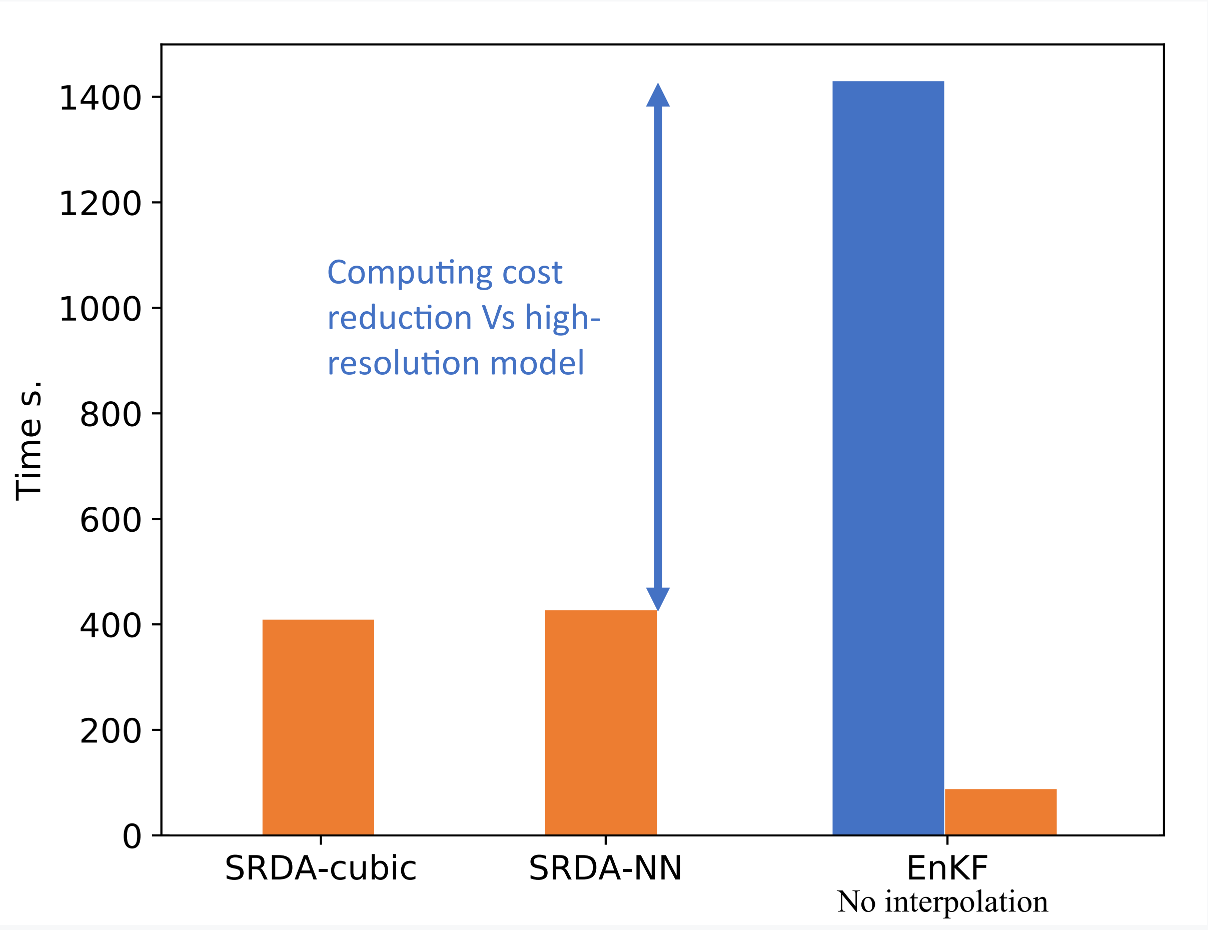
Performance of Super-Resolution data assimilation (SRDA)

NERSC

Using a high-resolution model

Using a low-resolution model





Barthélémy et al., 2022

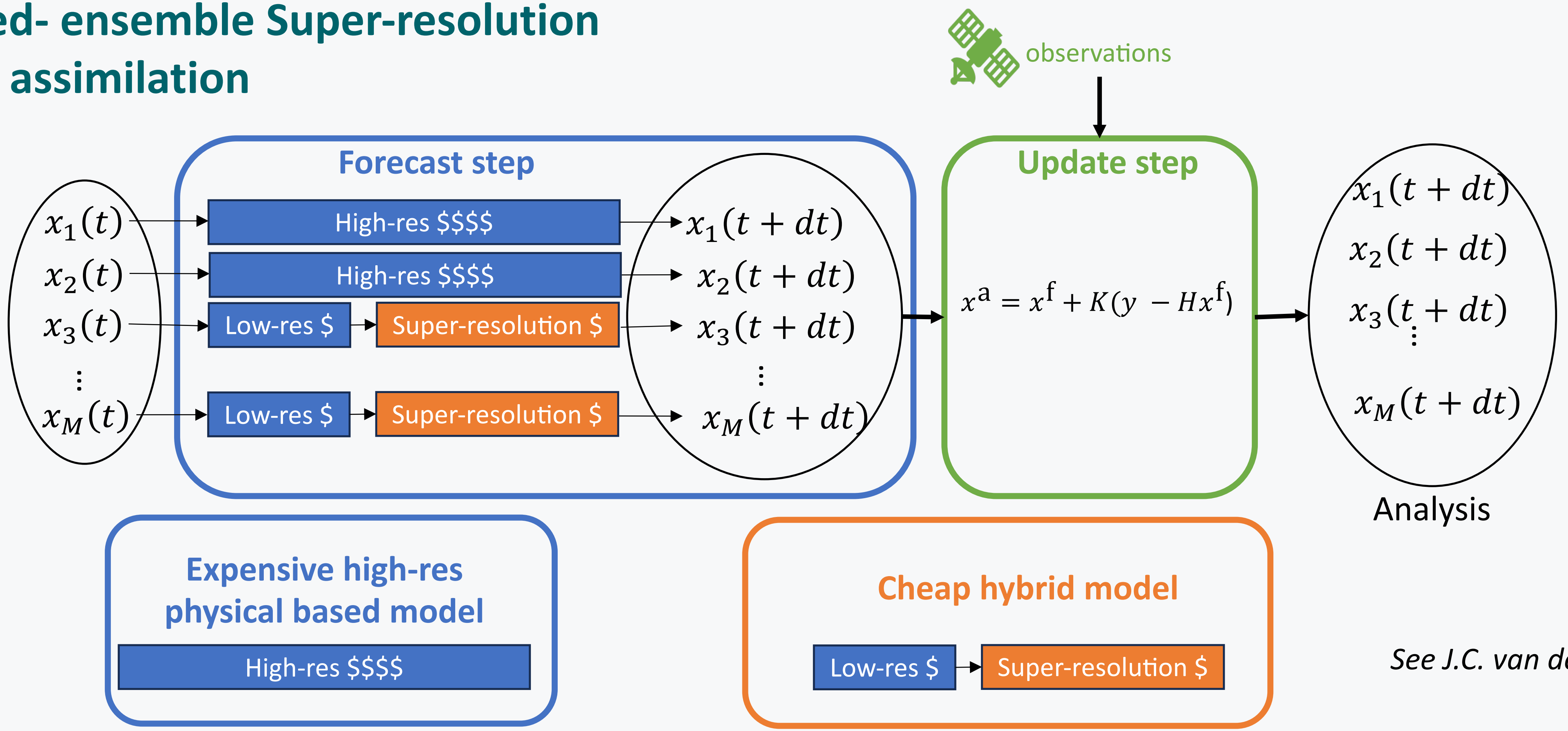
Full data-driven Vs hybrid models

NERSC

	Full data-driven	Hybrid model
Stability	**Can be unstable on long run**	**Generally stable**

One step further

Mixed- ensemble Super-resolution data assimilation



See J.C. van der Voort's talk

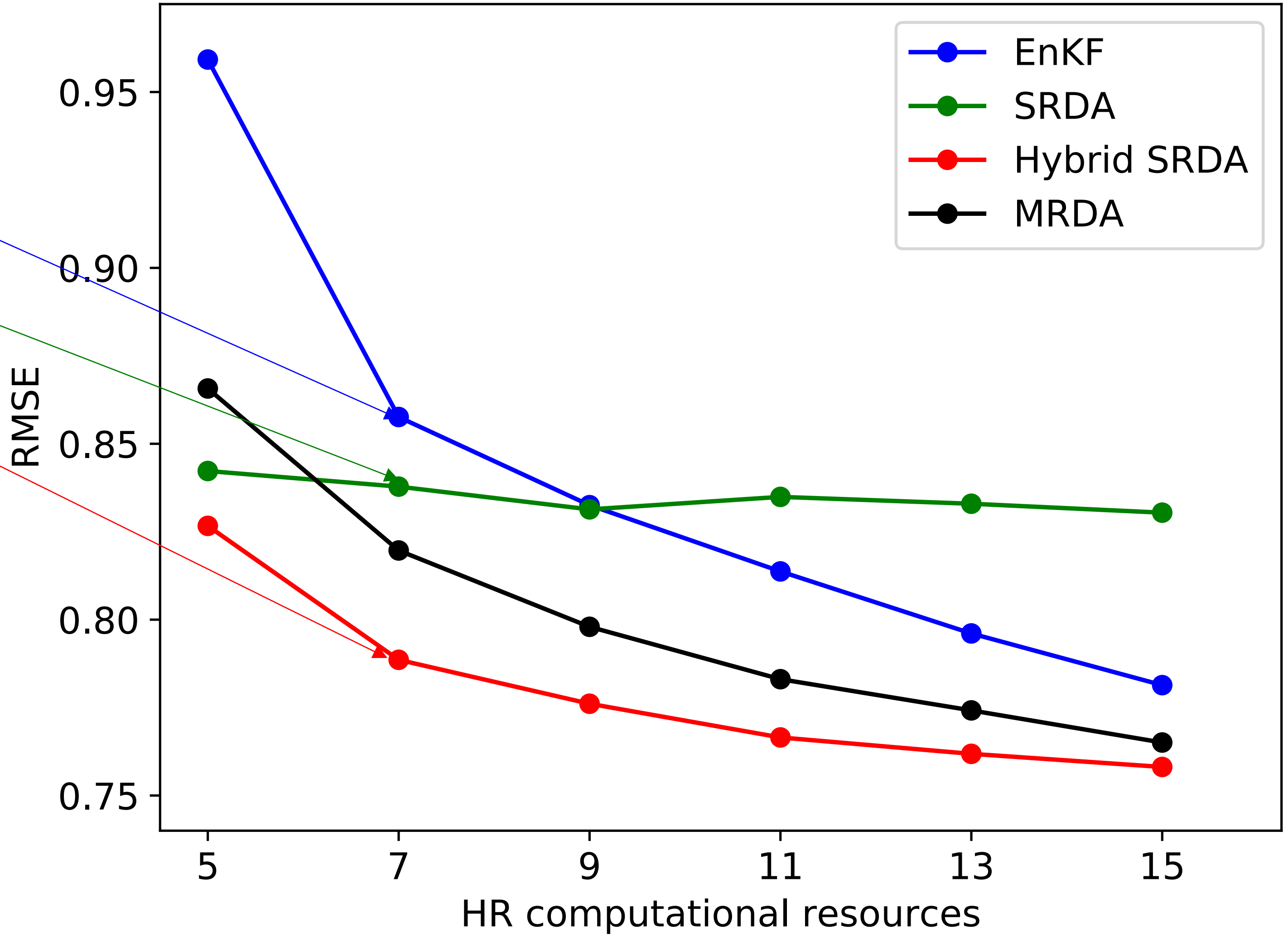
Barthélémy et al., 2024 (preprint)

Results

7 physics-based members

56 ML hybrid-based members

43 members
3 physics-based & 40 ML



For the same cost, it is beneficial to mix physics-based and data-driven forecasts

Barthélémy et al., 2024 (preprint)

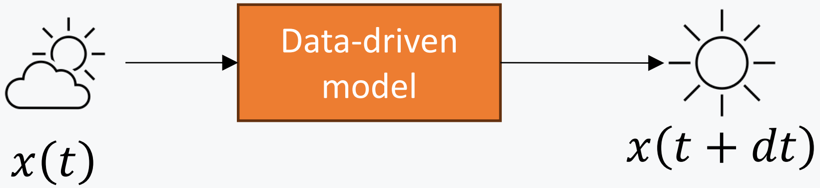
How to overcome this underestimation of spread (i.e. blurring)?

Idea #1
Hybrid models

Idea #2
Generative models

Two types of data-driven models

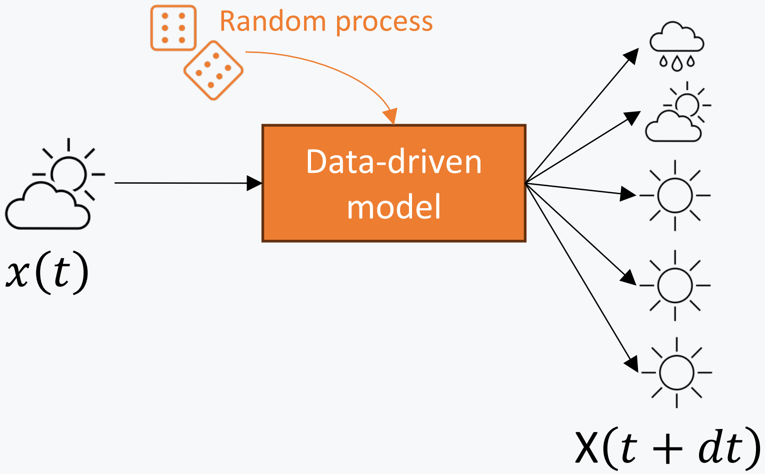
Deterministic model



One input → One output

- 🙂 Training is stable and converges quickly
- 🙂 Minimize the root-mean-square error
- ☹ Does not provide an uncertainty estimate
- ☹ Blurred outputs

Generative/stochastic model

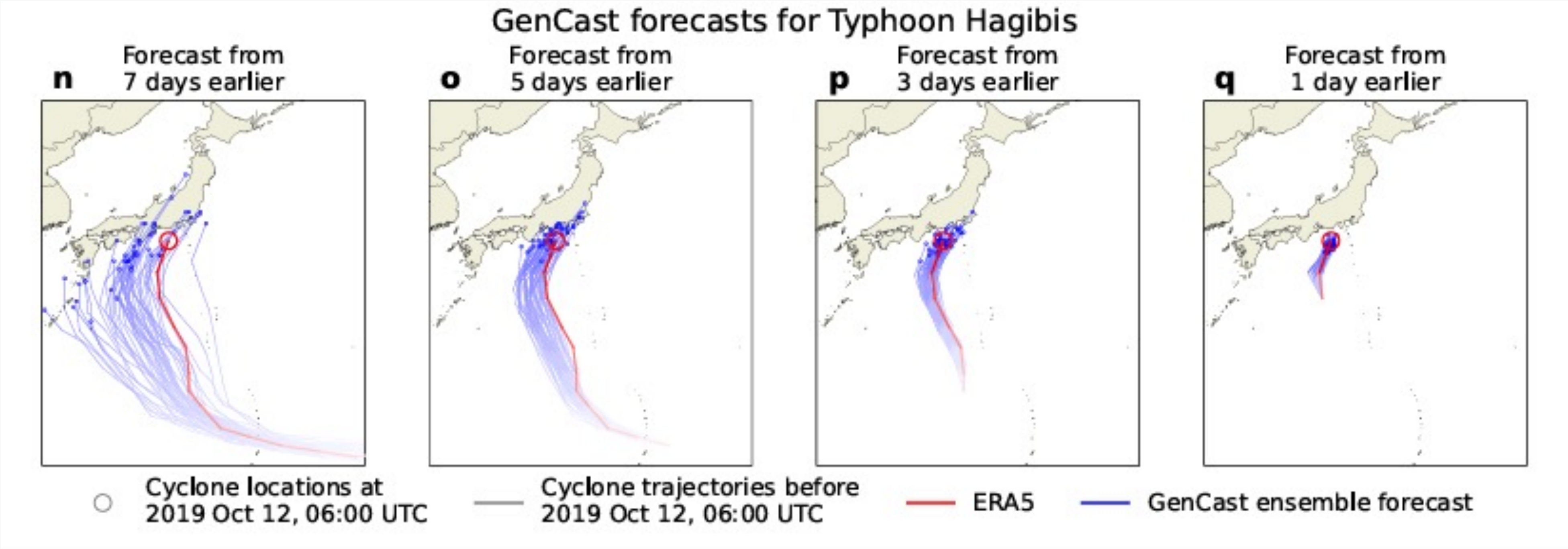


One input → A ensemble/distribution

- ☹ Training is more challenging and needs more data
- ☹ Validation metrics are multiple
- 🙂 Provides an uncertainty estimate
- 🙂 Realistic outputs

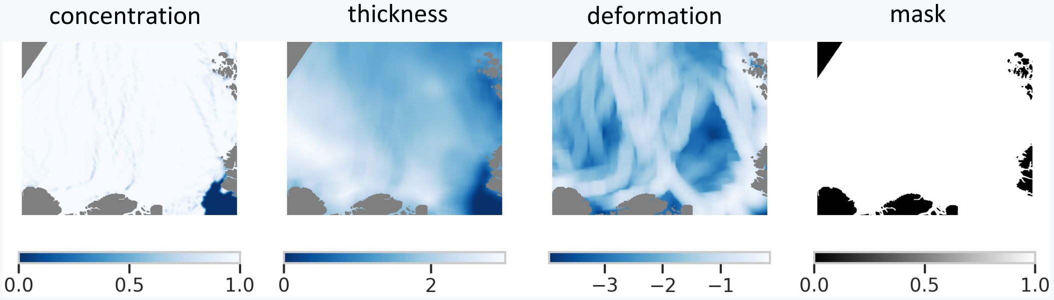
Extreme events

Generative model

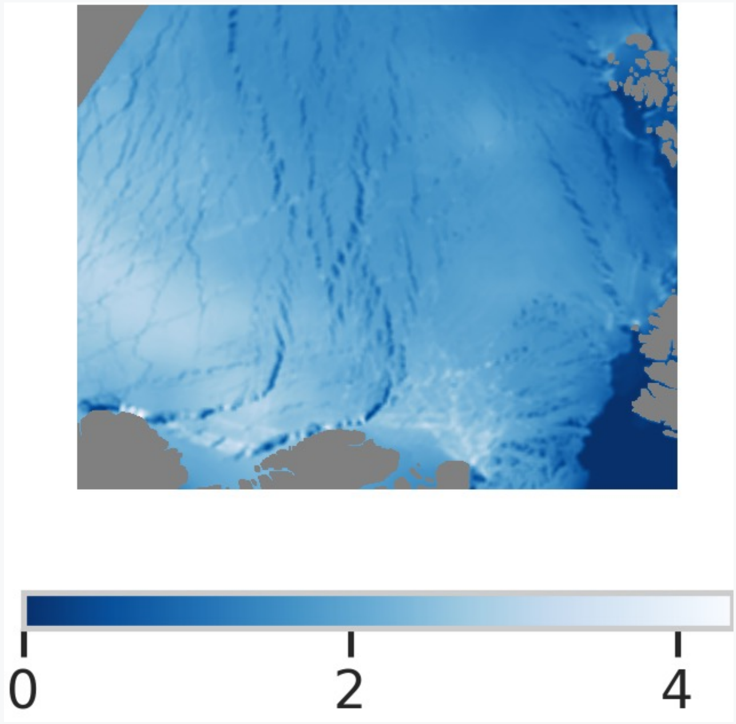


Gencast, Price et al., 2024

SuperIce

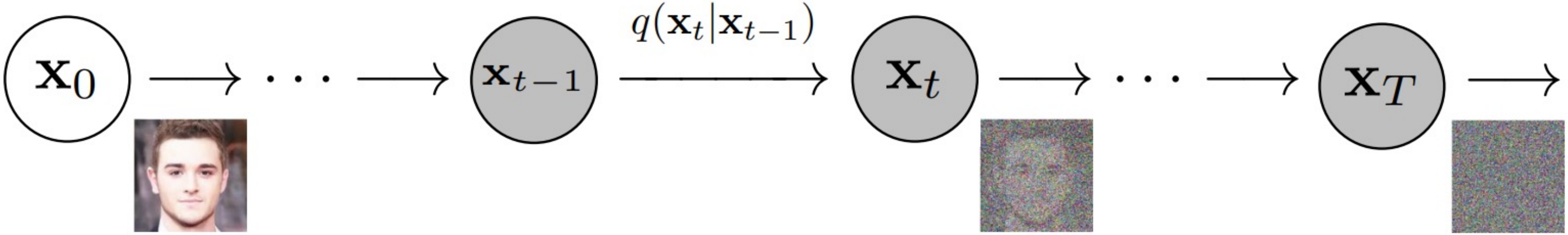


Generative diffusion model

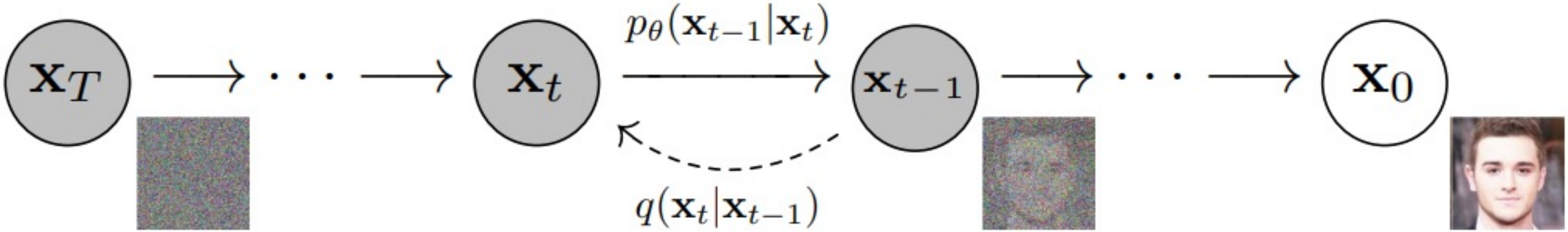


Principle of diffusion models

Noising procedure



Denoising procedure

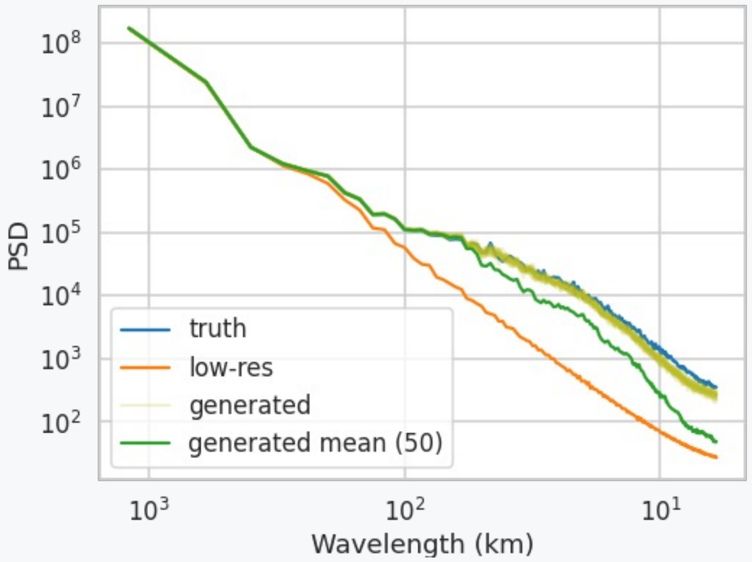


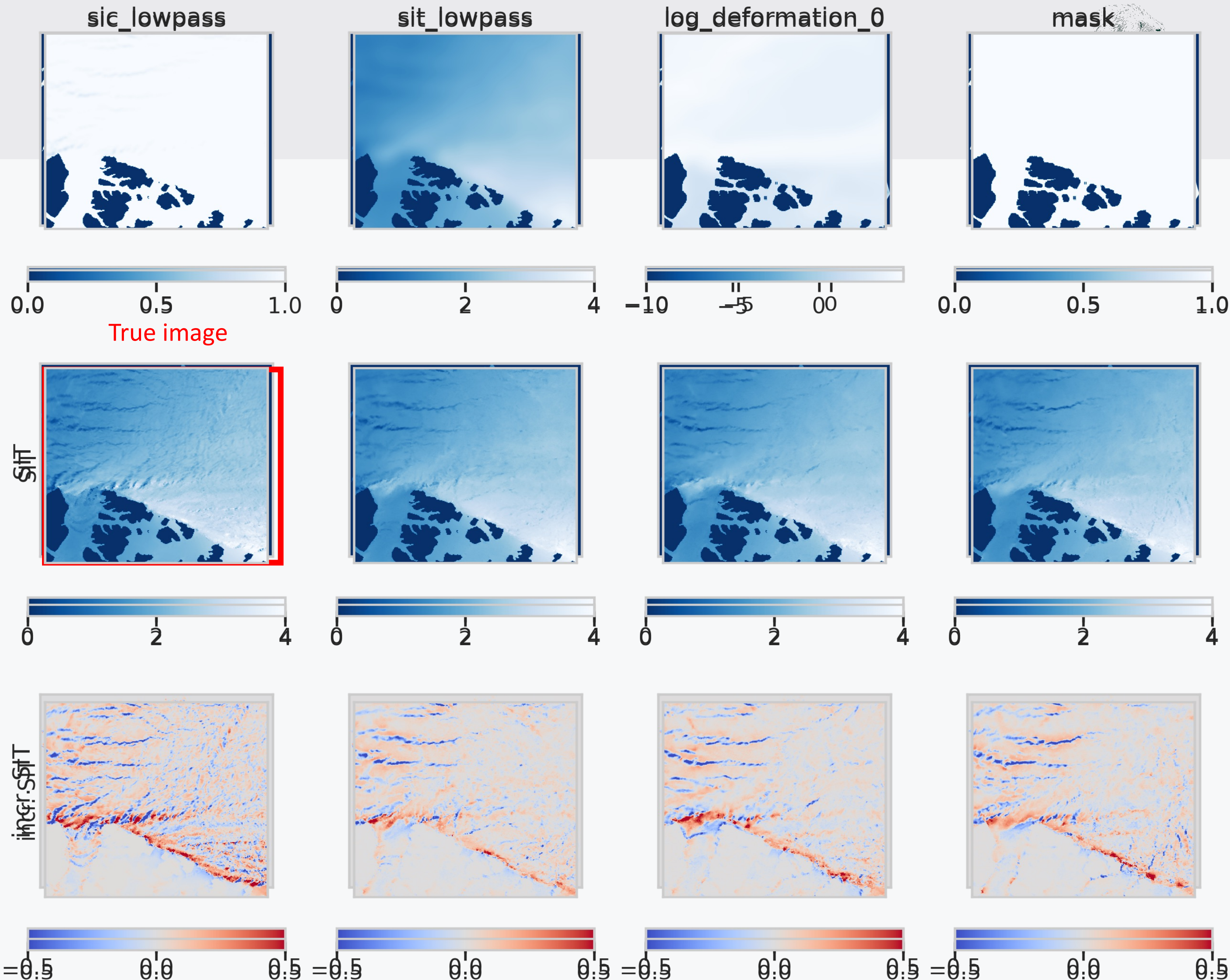
Super-resolution

Increase the resolution of SIT
(Sea Ice Thickness)

Use of diffusion model
Example 26 Jan 2016

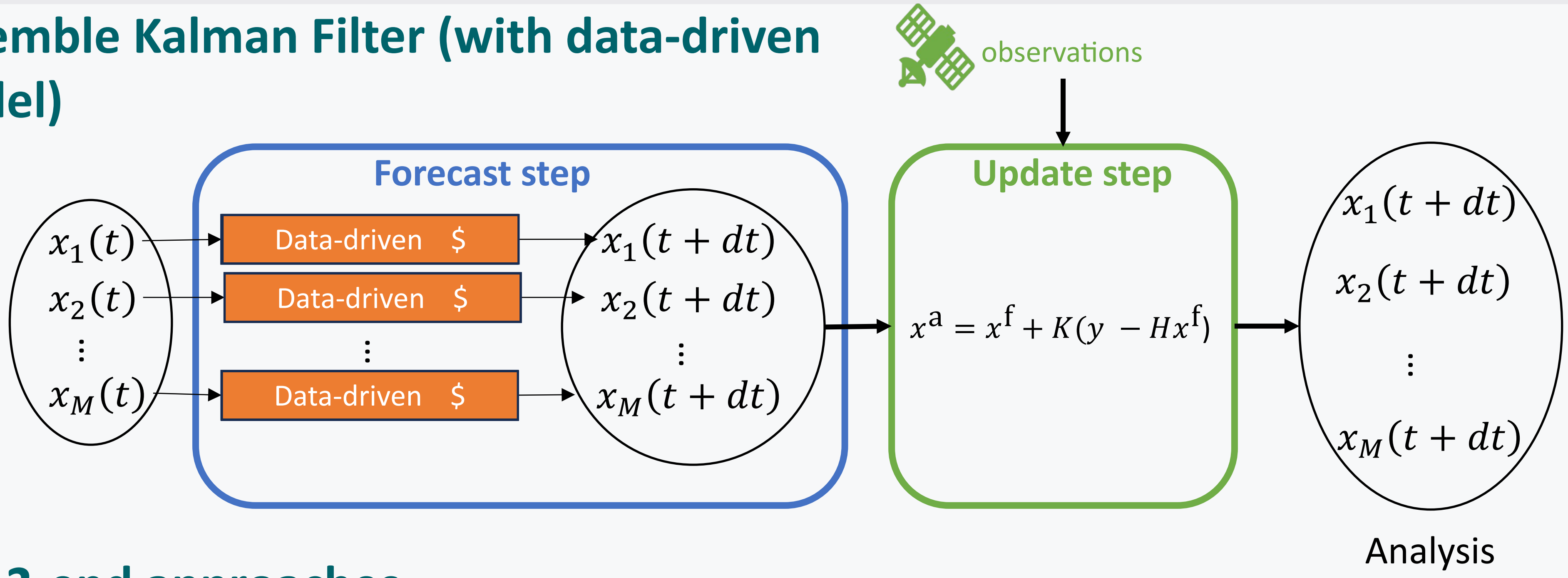
Preservation of most of the small scales



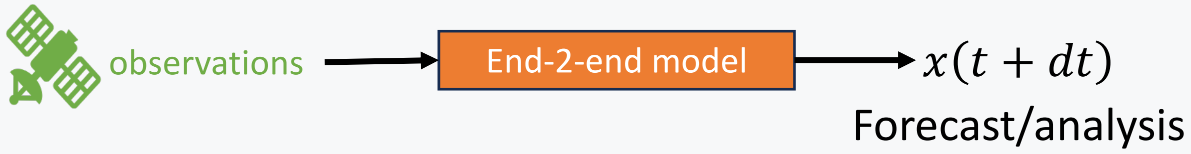


End to end?

Ensemble Kalman Filter (with data-driven model)



End-2-end approaches



End-2-end?

- ✓ Limited to well-observed variables
- ✓ Limited to the observed spatial and temporal resolution
- ✓ Is there enough data to learn the multivariate complex relationship without the constraint of a physical model?

Examples of End-2-end approaches

- ✓ Specific variables of interests
  - ✓ e.g. precipitation: *MetNet, Sønderby et al. 2020*
  - ✓ Sea ice: Kvanum et al., 2023
- ✓ First attempts for weather forecast
  - ✓ e.g. Aardvark Weather, *Vaughan et al. 2024*
  - ✓ *Lessig et al. 2024*

Take home message

NERSC

- **Machine learning** can help producing **data-driven models** that can be a computing-efficient alternative of **physical models** in a data assimilation framework
- **Deterministic data-driven models** can display problems of **stability, accuracy, and blurring** that could limit their use.
- **Hybrid-model and/or Generative models** are a promising way to overcome the problems of **deterministic data-driven models**
- For specific cases, **end-2-end approaches** can also be a option.

Julien.brajard@nersc.no

References

- Durand, 2024 Four-dimensional variational data assimilation with a sea-ice thickness emulator, https://zenodo.org/records/11636139
- Lam, 2023, GraphCast: Learning skillful medium-range global weather forecasting, https://arxiv.org/abs/2212.12794
- Rasp, 2024, The second revolution in numerical weather prediction and its implications for data assimilation, https://www.youtube.com/watch?v=CoiVfwJU4TY
- Xiao, 2023, FengWu-4DVar: Coupling the Data-driven Weather Forecasting Model with 4D Variational Assimilation, https://arxiv.org/abs/2312.12455
- Ying, 2024, NEDAS package, https://github.com/nansencenter/NEDAS/tree/develop
- Barthélémy, S., Brajard, J., Bertino, L. and Counillon, F., 2022. Super-resolution data assimilation. Ocean Dynamics, 72(8), pp.661-678. https://link.springer.com/article/10.1007/s10236-022-01523-x
- Kochov et al. 2023, Neural General Circulation Models for Weather and Climate, https://arxiv.org/abs/2311.07222
- Barthélémy, S., Counillon, F., Brajard, J. and Bertino, L., 2024. Hybrid covariance super-resolution data assimilation. https://www.researchsquare.com/article/rs-4454996/latest
- Price et al. 2023, Gencast: Diffusion-based ensemble forecasting for medium-range weather, https://arxiv.org/abs/2312.15796
- Brajard et al. 2024, Super-resolution of satellite observations of sea ice thickness using diffusion models and physical modeling., https://nansencenter.github.io/superice-nersc/assets/slides_and_posters/PICO_superice_2-Belgrade.pdf