UQ in weather forecasting :
From physics-based modeling to data-driven modeling

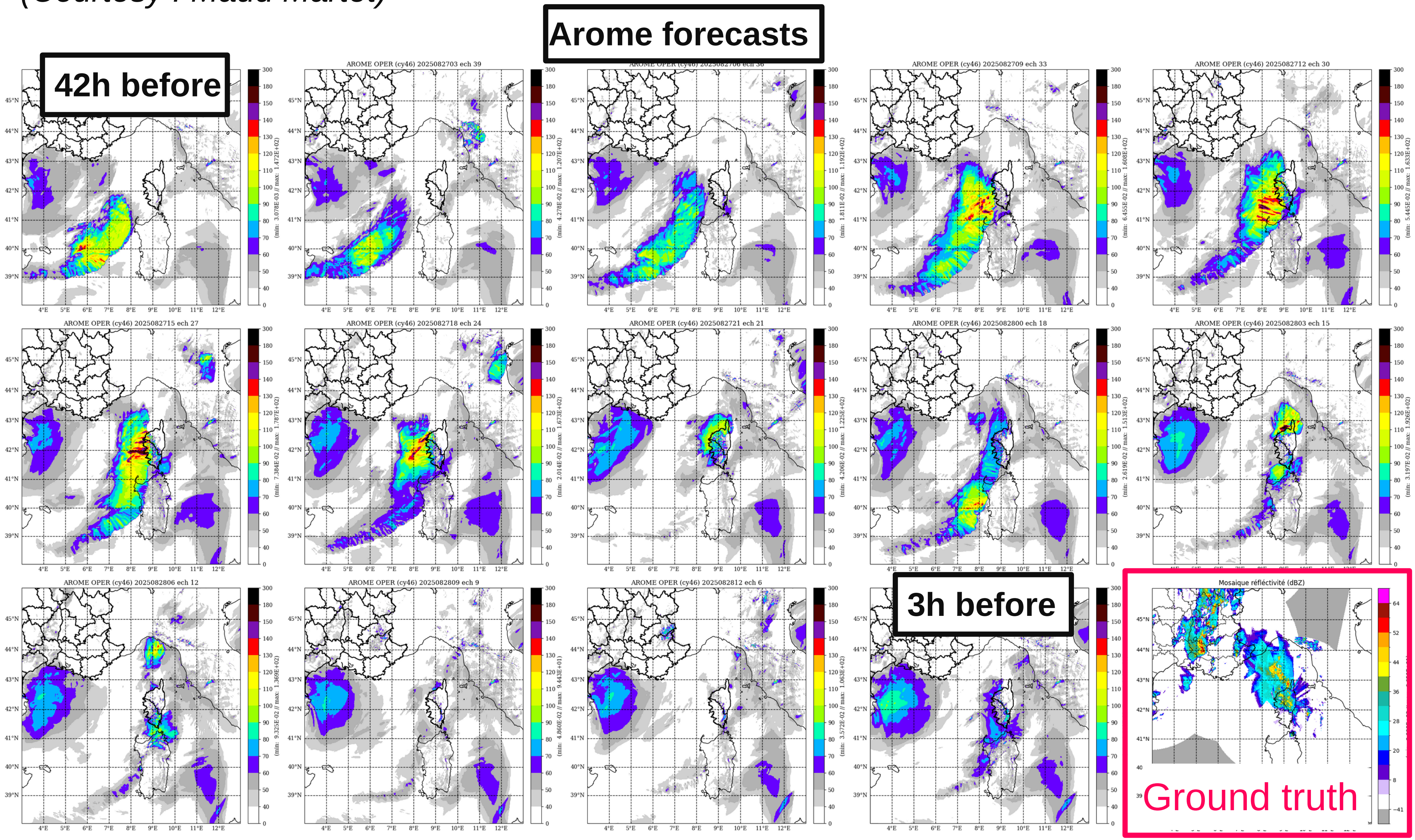
Laure Raynaud, Météo-France
13 November 2025

Outline

- Why do we need UQ in weather forecasting ?
- Main approaches for UQ in current operational systems
- UQ and machine learning
  - State-of-the-art, early results and challenges

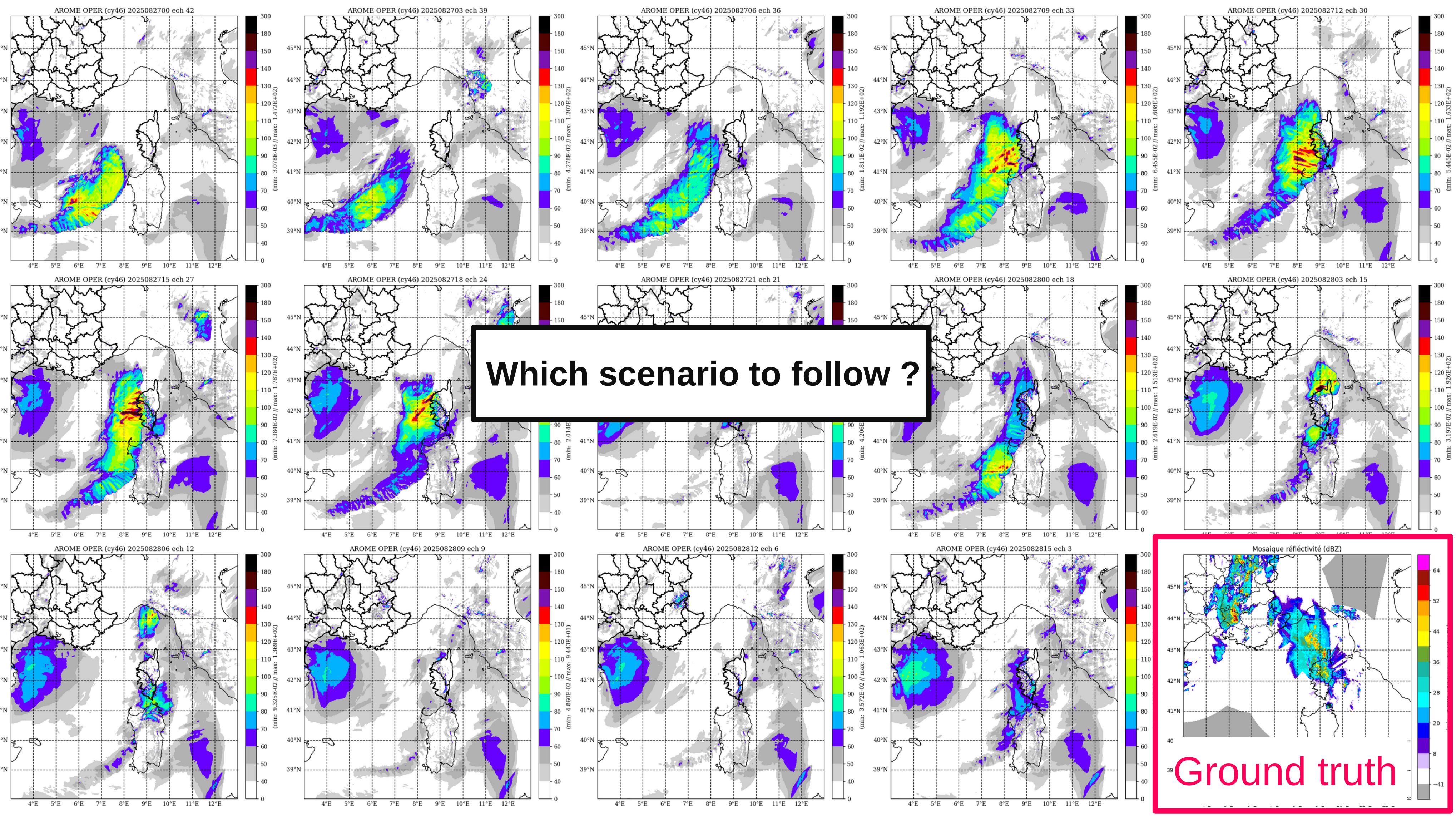
Back to the roots : why do we need UQ ?

How successive forecasts anticipate severe thunderstorms ?
(Courtesy : Maud Martet)



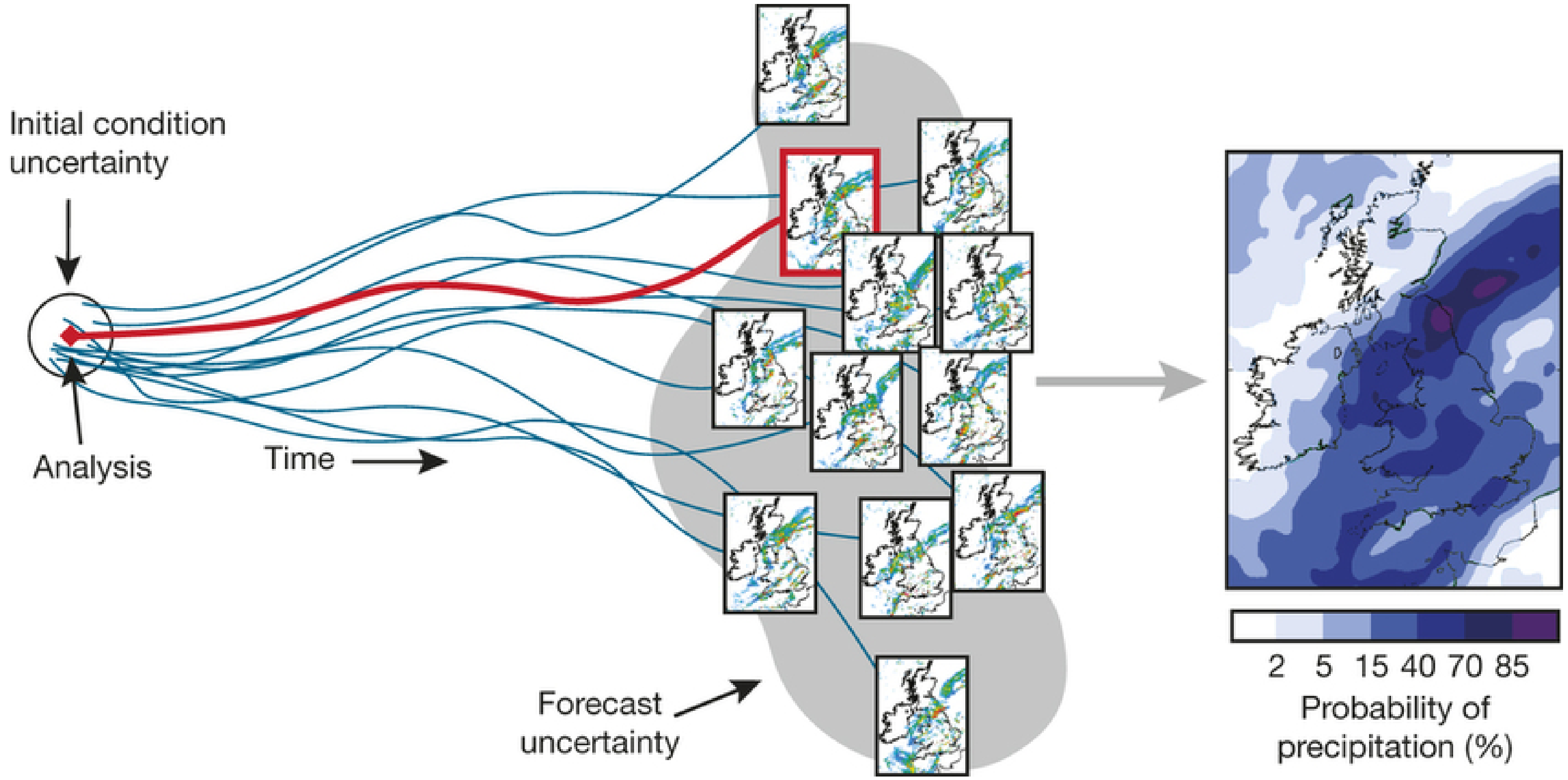
Why do we need UQ ?

Despite the increasing accuracy of weather forecasts, **uncertainty remains** and must be accounted for.



How to handle UQ in practice ?

➢ From deterministic to probabilistic forecast : Monte Carlo estimation of the forecast probability distribution, also known as **Ensemble Forecasting**



How to build reliable ensembles ?

Identify sources of uncertainty

- Initial conditions
- Model formulation
- Couplings (lateral boundary conditions, other earth system components)

Design perturbation strategies

- To properly sample uncertainty sources
- Computationally efficient

Ensemble size

- Sufficiently large to get a reliable pdf estimate
- But strongly constrained by computational resources

State-of-the-art for ensemble perturbations

Initial uncertainty

- Ensemble data assimilation (talk by M. Bocquet)
- Singular vectors

Model uncertainty

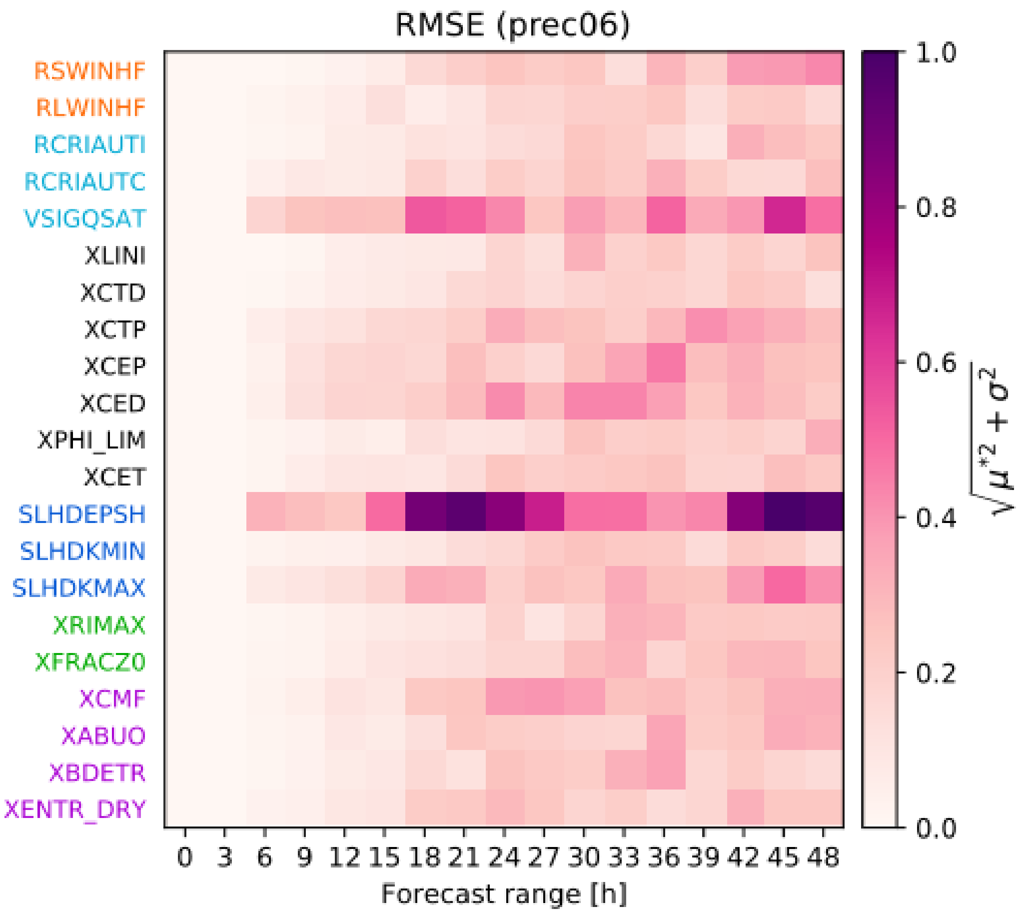
- Still challenging to fully characterize and represent
- A diversity of approaches to be used alone or in combination
  - post-hoc : inflation, perturbation of physics tendencies
  - physics/dynamics-based : multi models, multi physics, stochastic parametrizations, **parameter perturbations**

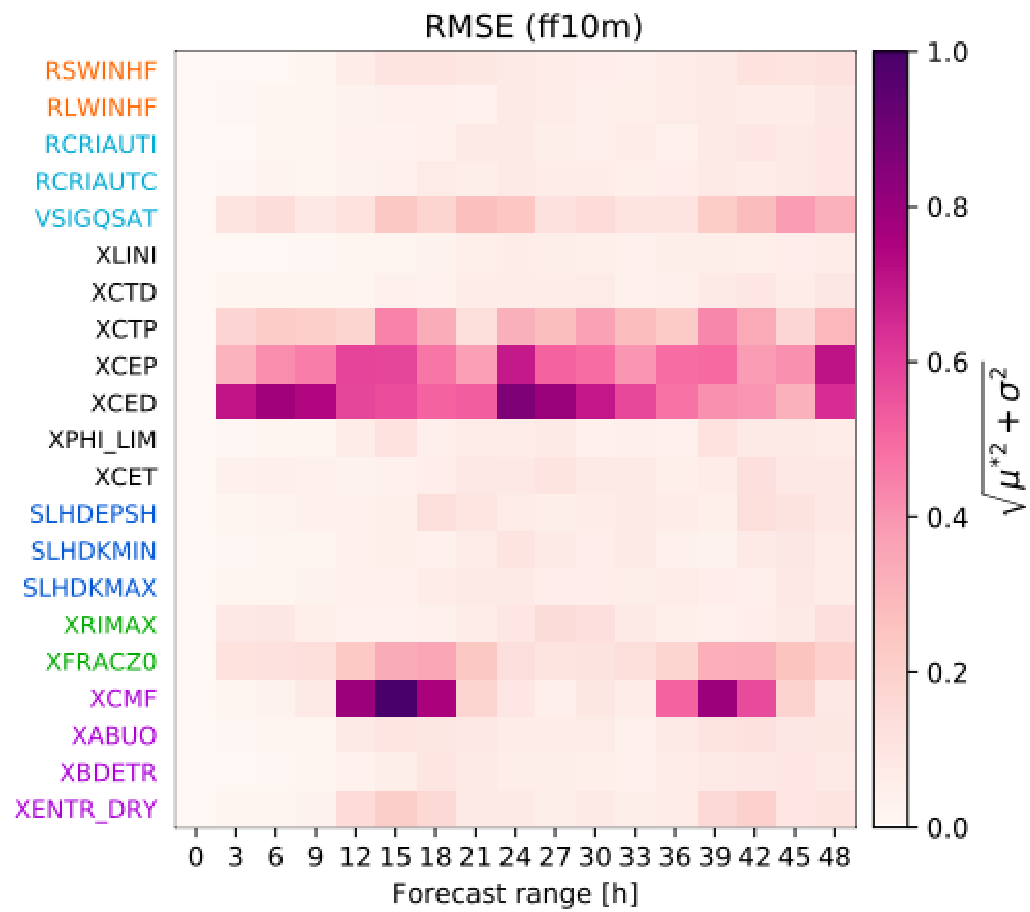
Ensemble size

- 10 to 50 members in general

Understanding model uncertainty : sensitivity of Arome forecasts to physics parameters

Goal : Identify a set of key parameters to perturb based on sensitivity analysis





Wimmer et al., 2021

Examples of state-of-the-art EPS : Météo-France operational systems

Arpège-EPS (global)

- 5km resolution over France, 35 members
- Ensemble data assimilation
- Singular vectors (to be abandoned very soon)
- Random parameter perturbations + multi-physics

Arome-France EPS (regional)

- 1,3km resolution, 25 members
- Ensemble data assimilation
- Perturbation of physics tendencies + random parameter perturbations
- Surface perturbations
- Ensemble lateral boundaries (from Arpège-EPS)

Next challenges

- **Uncertainty at very high resolution (hectometric scales)**
  - Higher computational costs
  - Faster loss of predictability at very fine scales
  - Many aspects of ensemble design at these scales are still unknown
    → How the nature of uncertainties might change at smaller spatial scales and how best to represent them ?
  - Only a few experimental trials so far (Hanley and Lean 2024)

- **Innovation in ensemble design seems to slow down**
  - Did we reach a plateau in our understanding and modeling of uncertainties ?

- **Uncertainty quantification and machine learning : the new paradigm ?**
  - ML to estimate errors (Farchi, Bocquet et al.)
  - ML to calibrate forecast distributions (Taillardat et al.)
  - *ML to produce larger ensembles*
  - *Ensembles with data-driven models*
    → Next part of the presentation

UQ and machine learning

- Can we use ML to boost current ensemble design ?
- How to revisit the challenge of UQ with the new data-driven weather forecasts ?

Can we use ML to boost ensemble design ?

- **Current ensemble performance is limited by**
  - Understanding and representation of uncertainty sources
  - Computational resources : <u>small ensembles</u>, cheap perturbation methods
- **Motivations for ML-based super-sampling**
  - Larger ensembles can improve uncertainty quantification
  - ML can generate members at a competitive computational cost
    - Examples : Li et al., 2024 ; Brochet et al. (2023, 2025)

Generative emulation of weather forecast ensembles with diffusion models

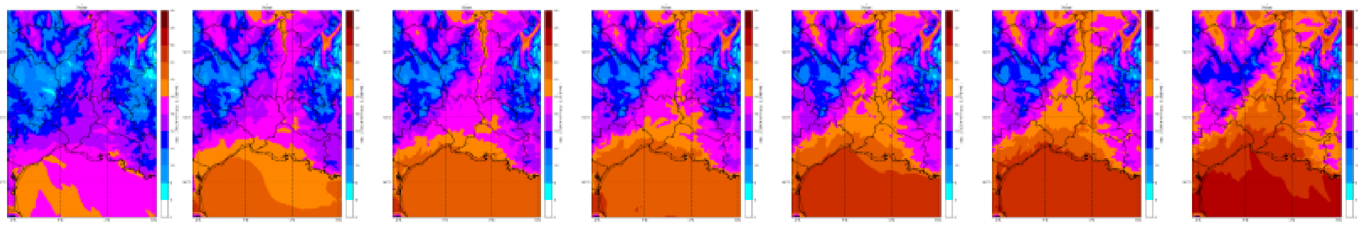
LIZAO LI, ROBERT CARVER, IGNACIO LOPEZ-GOMEZ, FEI SHA, AND JOHN ANDERSON Authors Info & Affiliations

Enriching Operational High-Resolution Ensemble Forecasts with StyleGAN-2

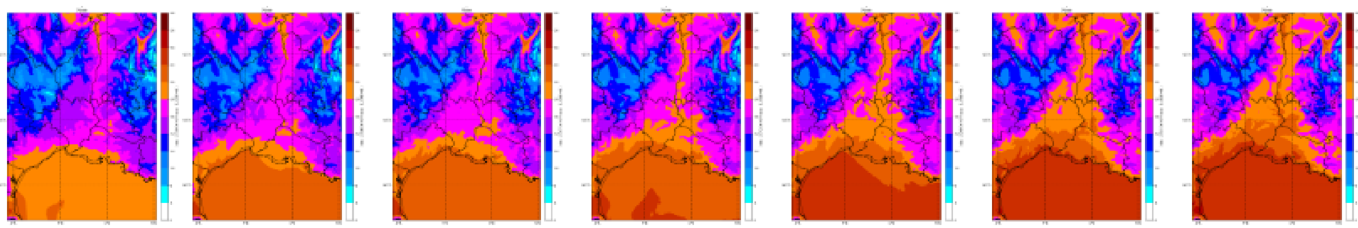
Clément Brochet,[a] Gabriel Moldovan,[a,b] Julien Rabault,[c] Cyril Regan,[d] and Laure Raynaud[a]

Motivation for larger ensembles : an example

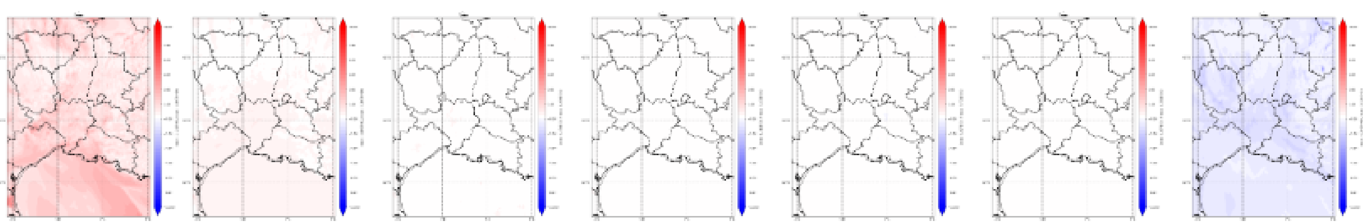
- 2m temperature percentiles 0, 10, 25, 50, 75, 90, 100.



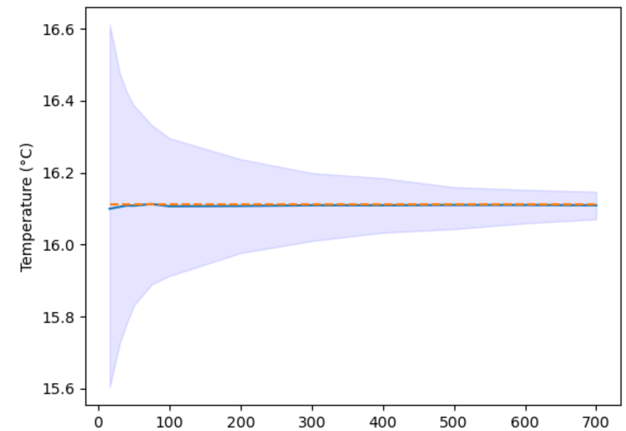
(a) 875mb

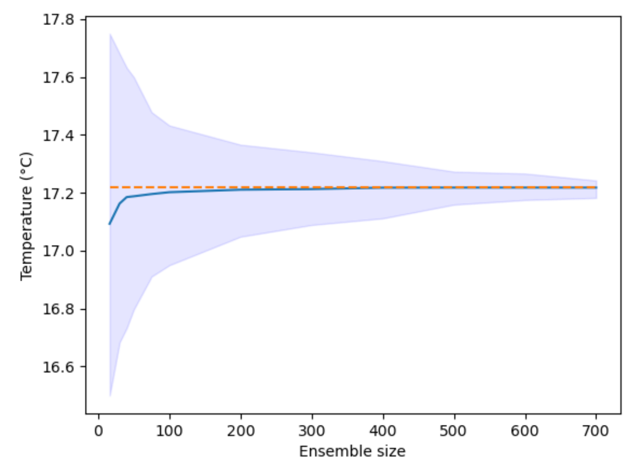


(b) 16mb



(c) Diff

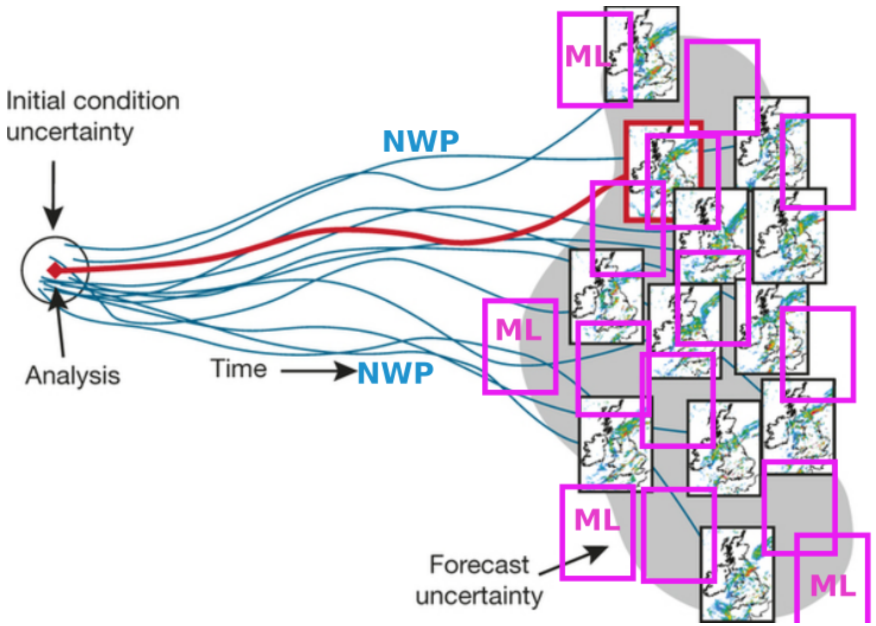




- Over/under-estimation of extreme percentiles + sampling errors

(Craig et al., 2022 ; Tempest et al., 2022)

ML to boost Arome-EPS : the concept

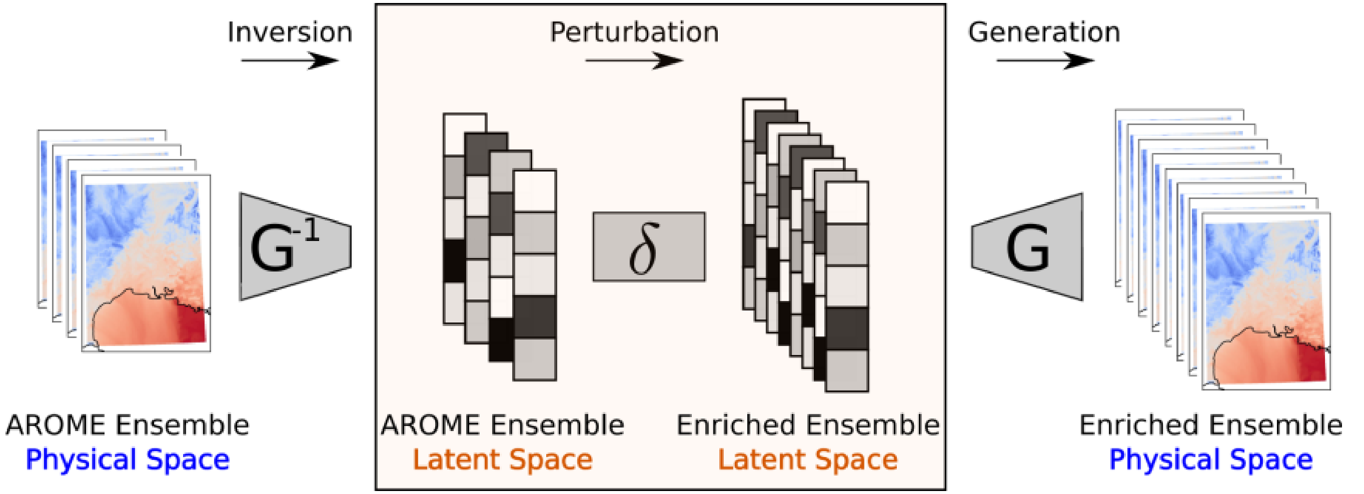


ML used for state estimation only (no propagation)

- ML members are tethered to a small NWP ensemble (16 Arome-EPS members)
- **Experimental setup**
  - Joint generation of 2-meter temperature and 10-meter wind
  - Training on ~ 1 year of Arome-EPS
- **Research questions**
  - Are the ML members physically-consistent ?
  - What is the added value of ML members ?
  - Can the ML ensemble emulate a very large NWP ensemble ?
  - Which ML algorithm is best suited to this task ?
  - Which type of uncertainty ML represents ?

ML to boost Arome-EPS

- **3 generative models to be compared**
  - Latent space image editing with **StyleGAN** (Brochet et al., 2025)

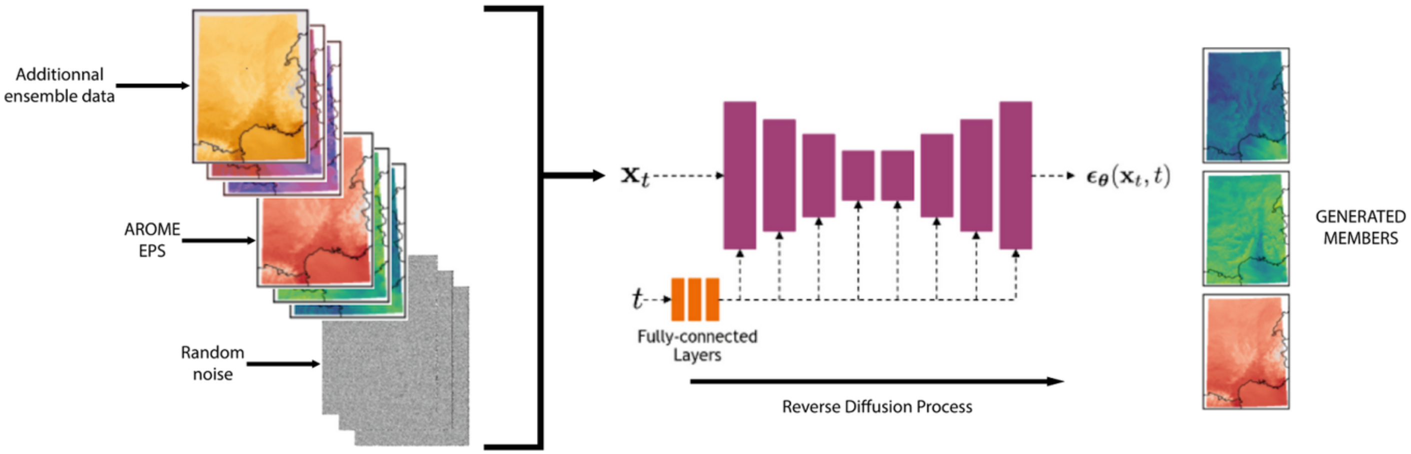


What to tune ?

➢ Inversion method
➢ Definition of perturbations

ML to boost Arome-EPS

- **3 generative models to be compared**
  - **Diffusion models** : (a) conditional diffusion, (b) image editing

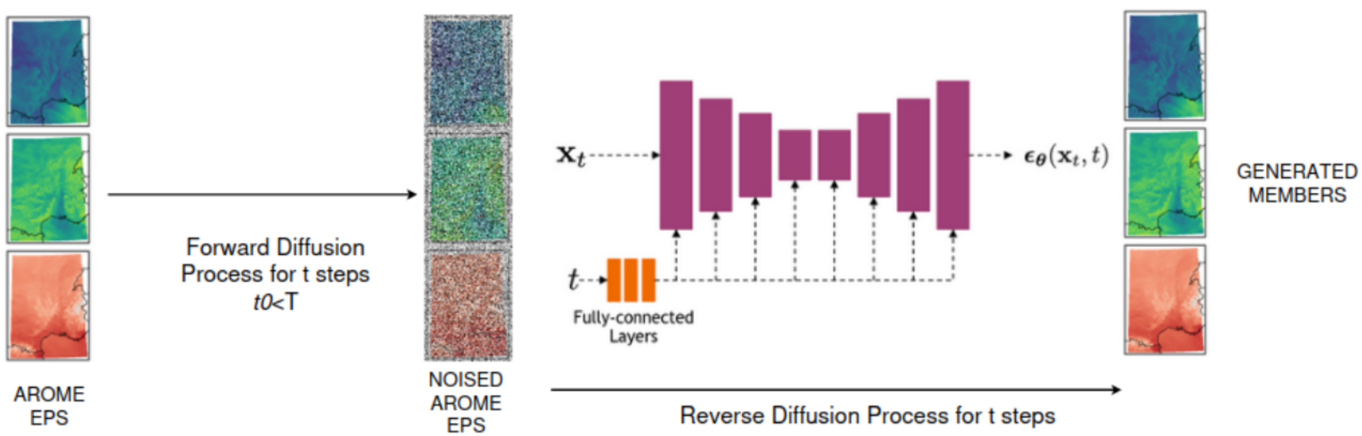


(a) Conditional diffusion

What to tune ?
- Conditioning information

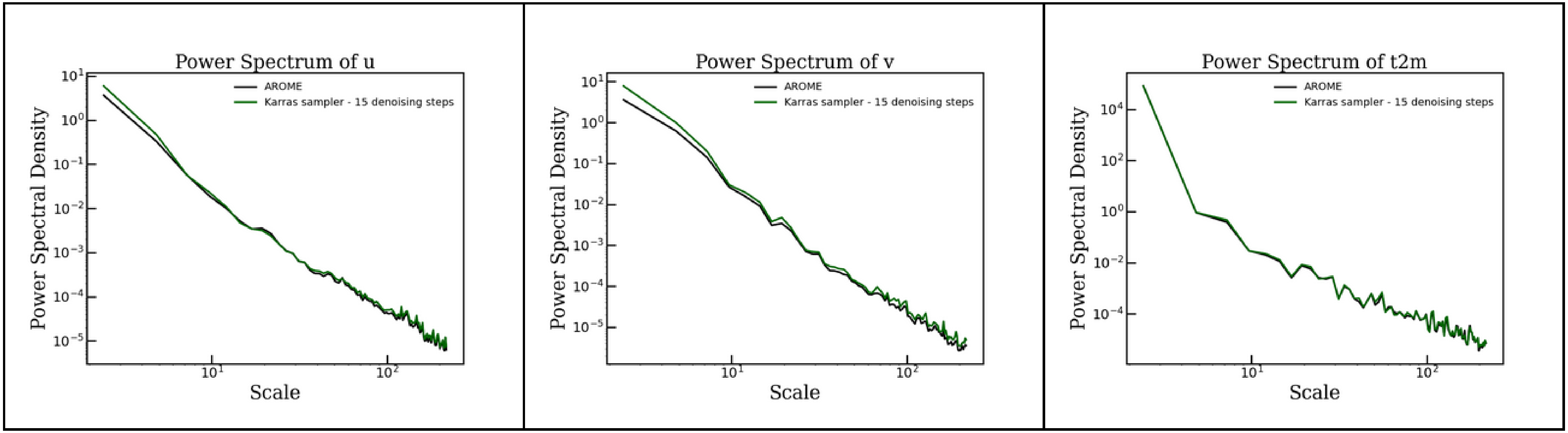
(b) SDEdit method (Meng et al.)

What to tune ?
- t_0



Physical consistency

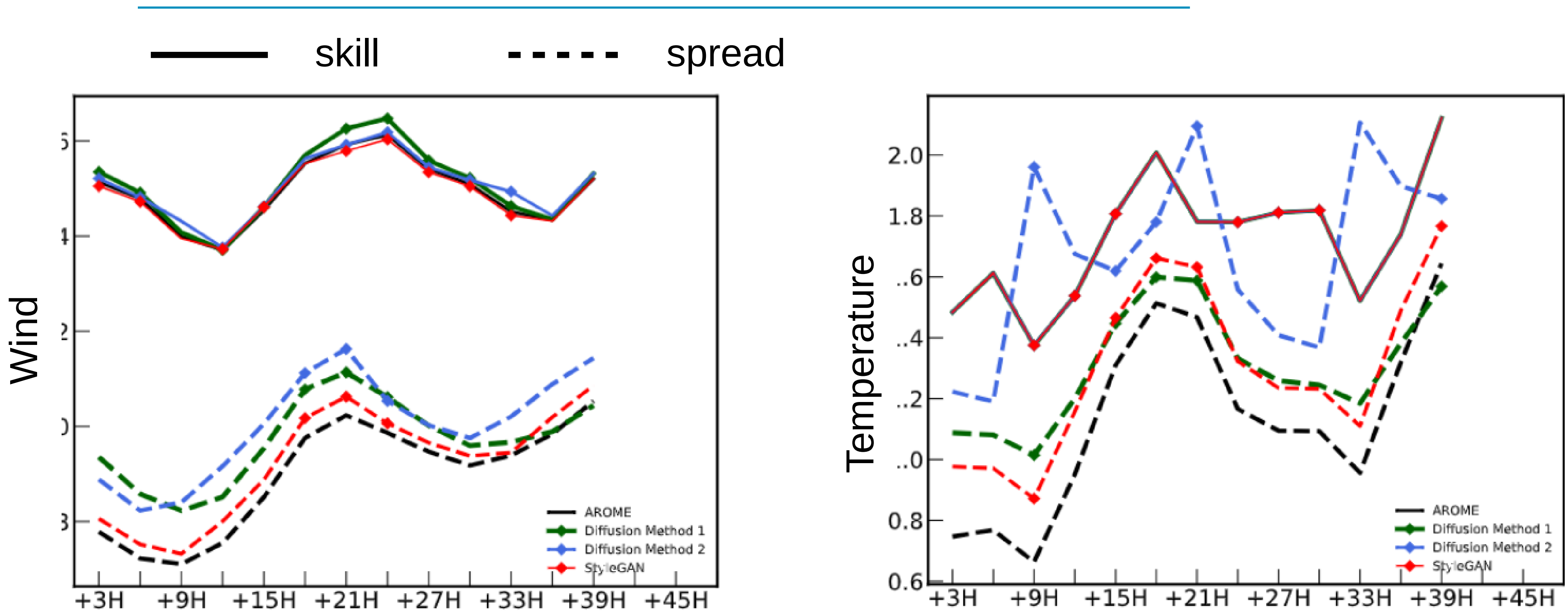
Arome Diffusion



Courtesy : M. Lame

- Diffusion is able to produce samples with the right amount of energy at all scales (similar results for StyleGAN)

Probabilistic performance : ML-boosted vs NWP EPS



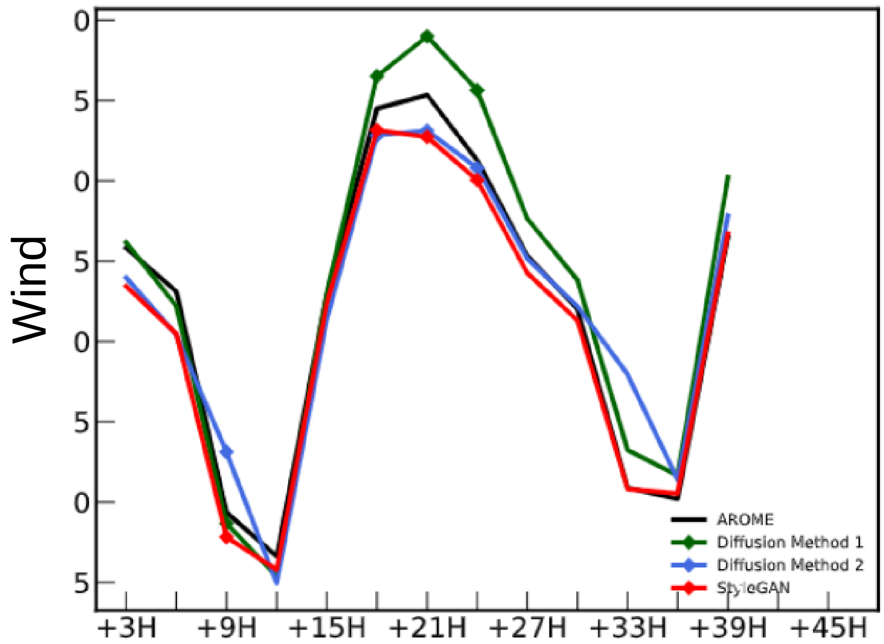
Arome StyleGAN Conditional diffusion SDEdit

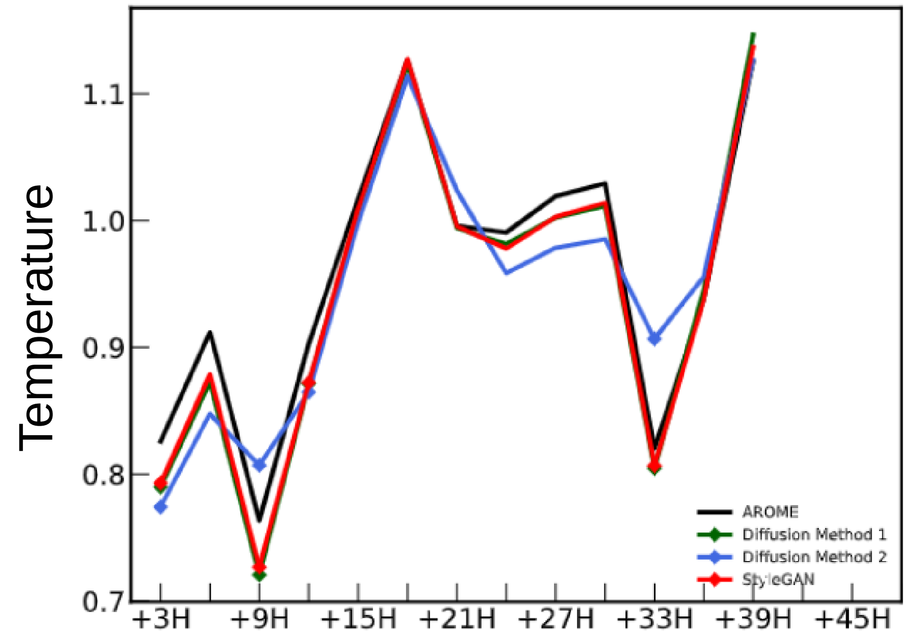
Courtesy : V. Sanchez

- ➢ Arome-EPS : 16 members, ML-boosted EPS : 64 members
- ➢ All ML models improve spread, SDEdit more dispersive

Probabilistic performance : ML-boosted vs NWP EPS

Fair CRPS





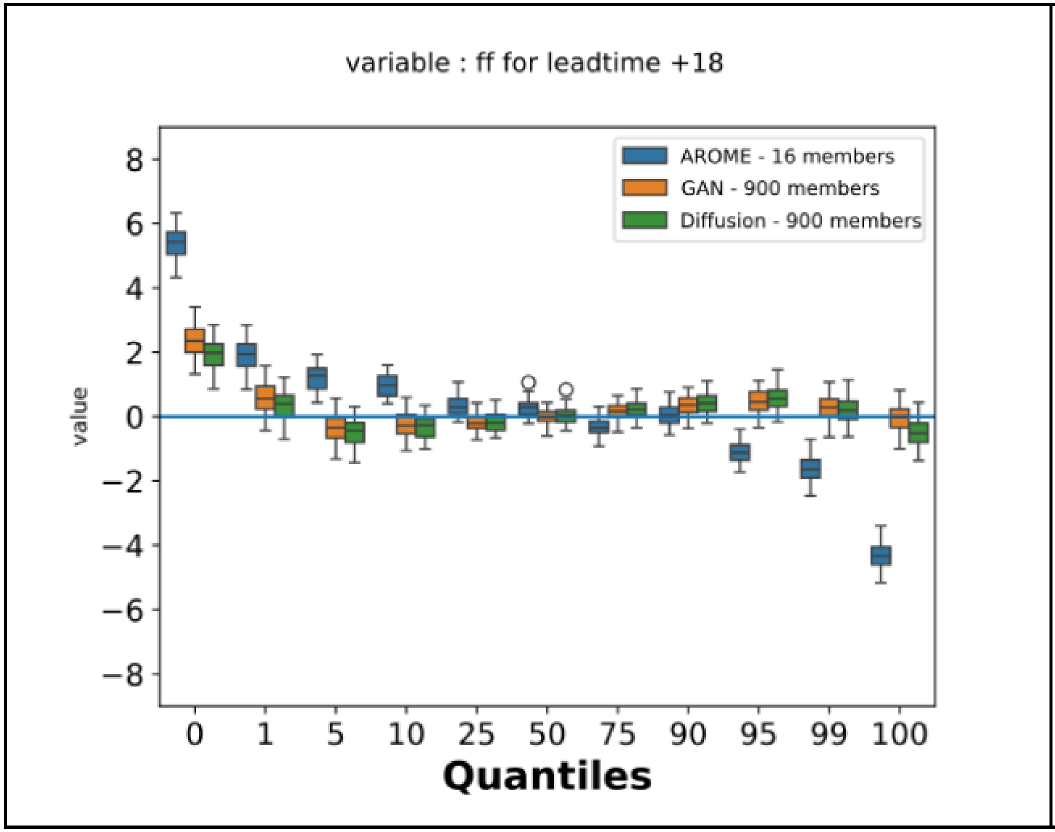
Arome StyleGAN Conditional diffusion SDEdit

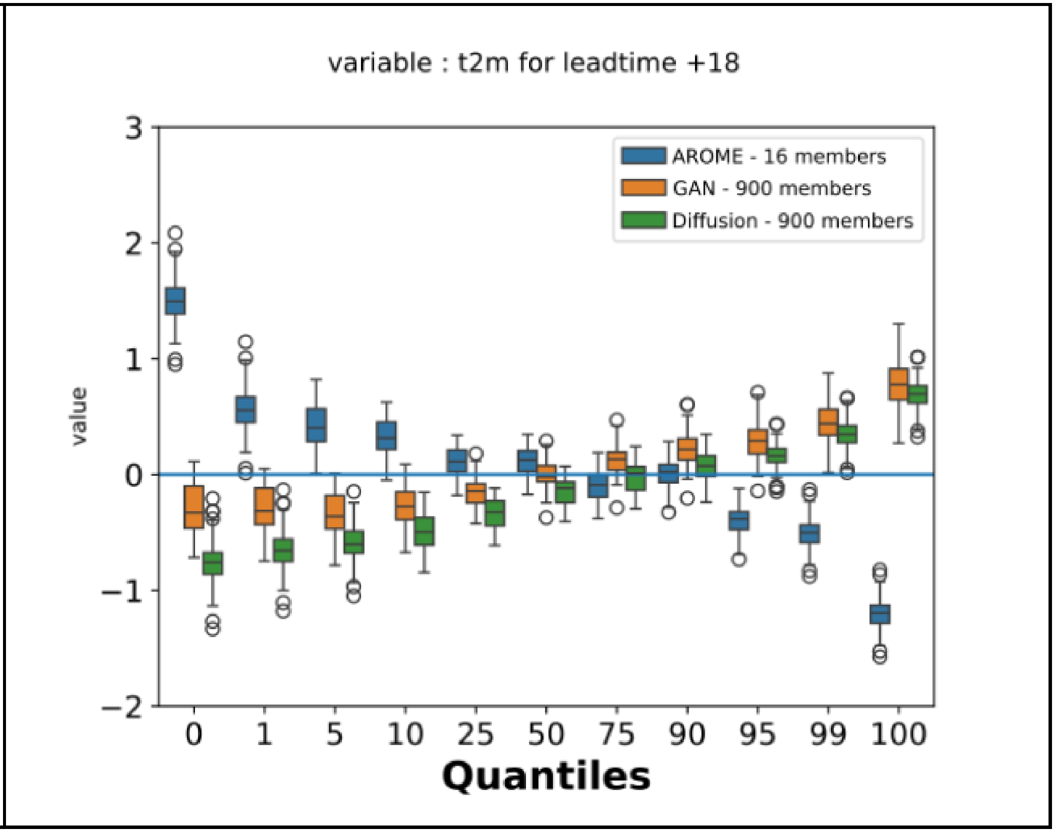
Courtesy : V. Sanchez

➢ Results are more mixed ...

ML-boosted EPS vs very large NWP EPS

StyleGAN and SDEdit 875 members vs Arome-EPS 875 and 16 members

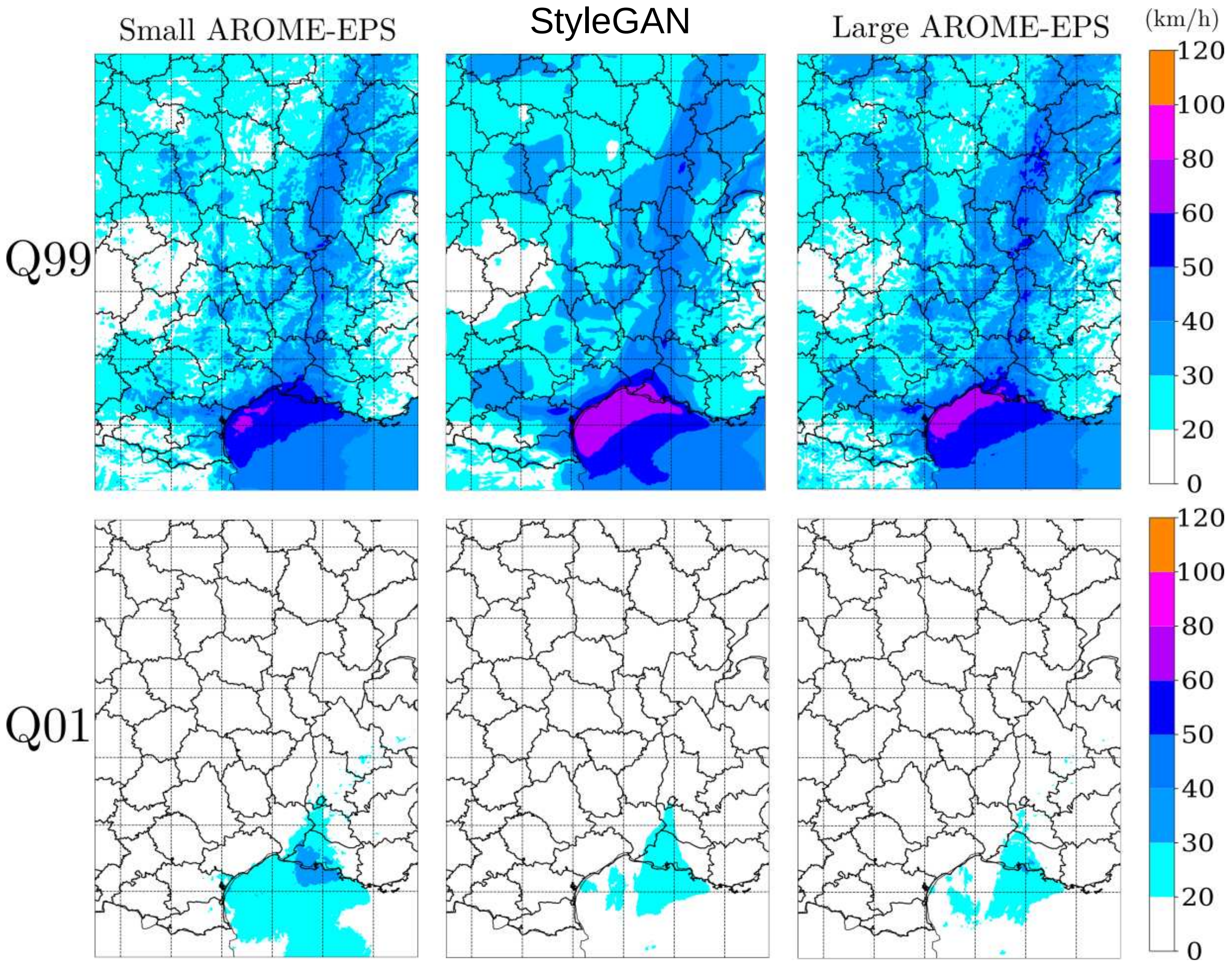




➢ Both ML methods reduce the underdispersion of Arome-EPS 16 and better match Arome-EPS 875

Courtesy : V. Sanchez

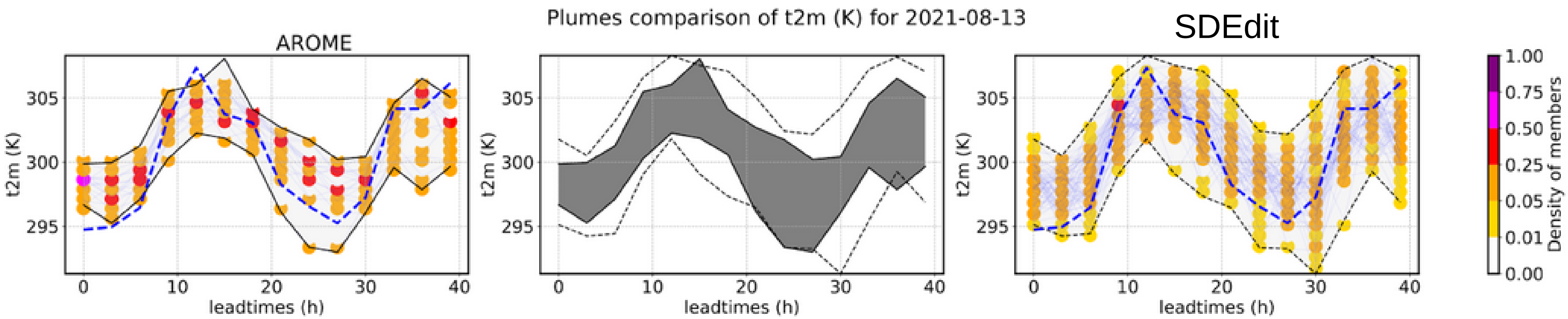
ML-boosted EPS vs very large NWP EPS



Brochet et al., 2025

Open questions and future works

- ML represents mis-represented sources of uncertainty but difficult to characterize which ones and how
- Scaling-up to more realistic setup (more variables, larger domain) still needs to be evaluated
- Operational usage/utility : need to better assess the relevance of ML-boosted EPS on extreme events
- Could also be applied to produce ensembles from deterministic forecasts, or to enrich initial conditions



Courtesy : V. Sanchez

Uncertainty quantification with data-driven model

- The rise of data-driven models is a **game changer** for operational forecasting
- It opens up a new way of producing very large ensembles at a low cost
- And new questions regarding the design of probabilistic forecasts

ECMWF unveils alpha version of new ML model

13 October 2023
The AIFS team

GenCast: Diffusion-based ensemble forecasting for medium-range weather

Ilan Price[*,1], Alvaro Sanchez-Gonzalez[*,1], Ferran Alet[1], Timo Ewalds[1], Andrew El-Kadi[2], Jacklynn Stott[1], Shakir Mohamed[1], Peter Battaglia[1], Remi Lam[1] and Matthew Willson[1]
[*]Equal contributions, [1]Google DeepMind, [2]Imperial College, London

AIFS OPERATIONAL

GraphCast: Learning skillful medium-range global weather forecasting

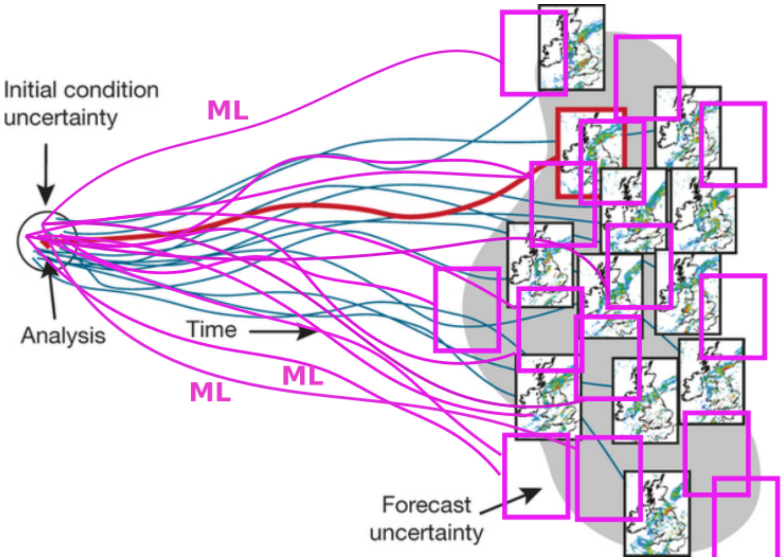
Remi Lam[*,1], Alvaro Sanchez-Gonzalez[*,1], Matthew Willson[*,1], Peter Wirnsberger[*,1], Meire Fortunato[*,1], …aton-Rosen[1], Weihua Hu[1], Alexander Merose[2], …lynn Stott[1], Alexander Pritzel[1], Shakir Mohamed[1] and

Pangu-Weather: A 3D High-Resolution System for Fast and Accurate Global Weather Forecast

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian, *Fellow, IEEE*

Uncertainty quantification with data-driven model

- **Different approaches for probabilistic forecasting**
  - A model that directly estimates the forecast distribution
  - A model that can produce members
    - Deterministic training + perturbations
    - Ensemble training
    - Generative deep learning



Not clear yet how these approaches compare !

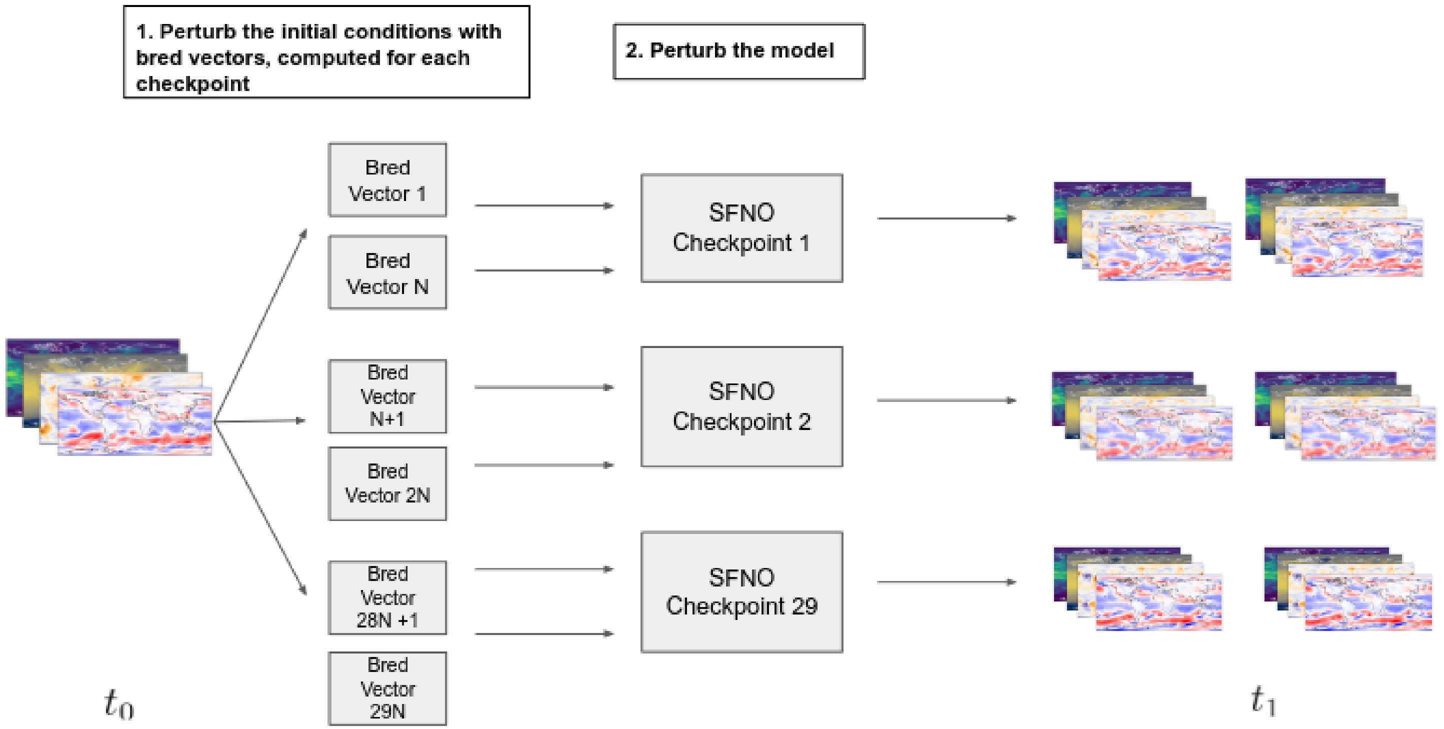
Uncertainty Quantification for Deep Learning

Peter Jan van Leeuwen[1,2], J. Christine Chiu[1], and C. Kevin Yang[1]

[1]Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA,
[2]Department of Meteorology, University of Reading, Reading, UK,

Deterministic model + perturbations

- Replicate the NWP EPS design with data-driven models

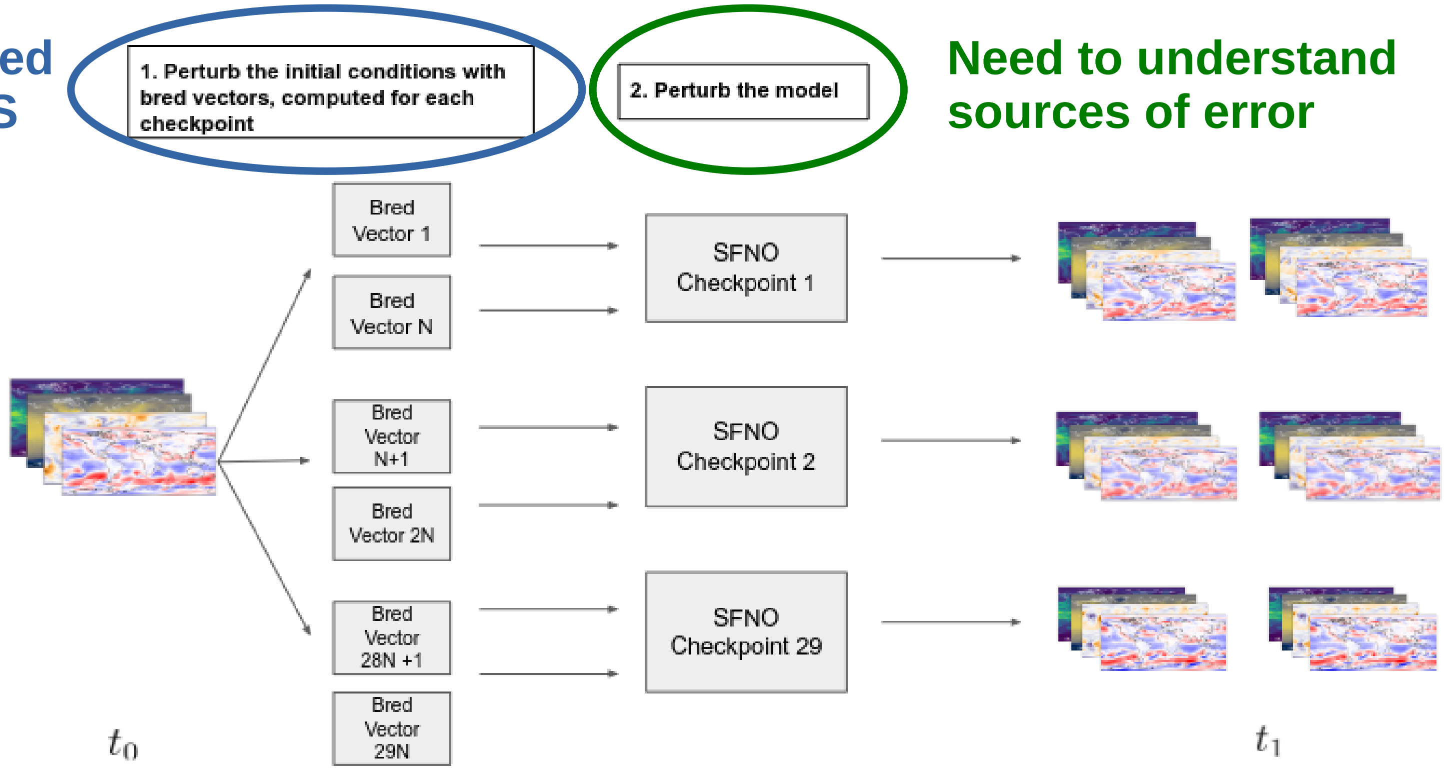


From Mahesh et al., 2024

Deterministic model + perturbations

- Replicate the EPS design with data-driven models

Could be reused from NWP EPS



Need to understand sources of error

From Mahesh et al., 2024

Challenges to define perturbations

- **Initial (aleatoric) uncertainty**

Unclear what is the sensitivity of ML models to initial conditions perturbations

Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect?

T. Selz[1] and G. C. Craig[1,2]

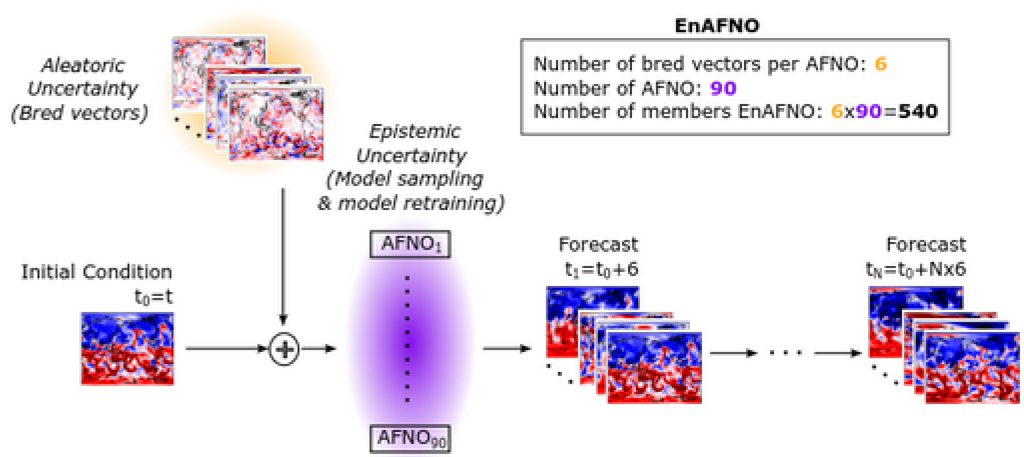
Uncertainty quantification for data-driven weather models

Christopher Bülte[a], Nina Horat[a], Julian Quinting[a] and Sebastian Lerch[a,b]

[a] *Karlsruhe Institute of Technology, Karlsruhe, Germany*

[b] *Heidelberg Institute for Theoretical Studies, Heidelberg, Germany*

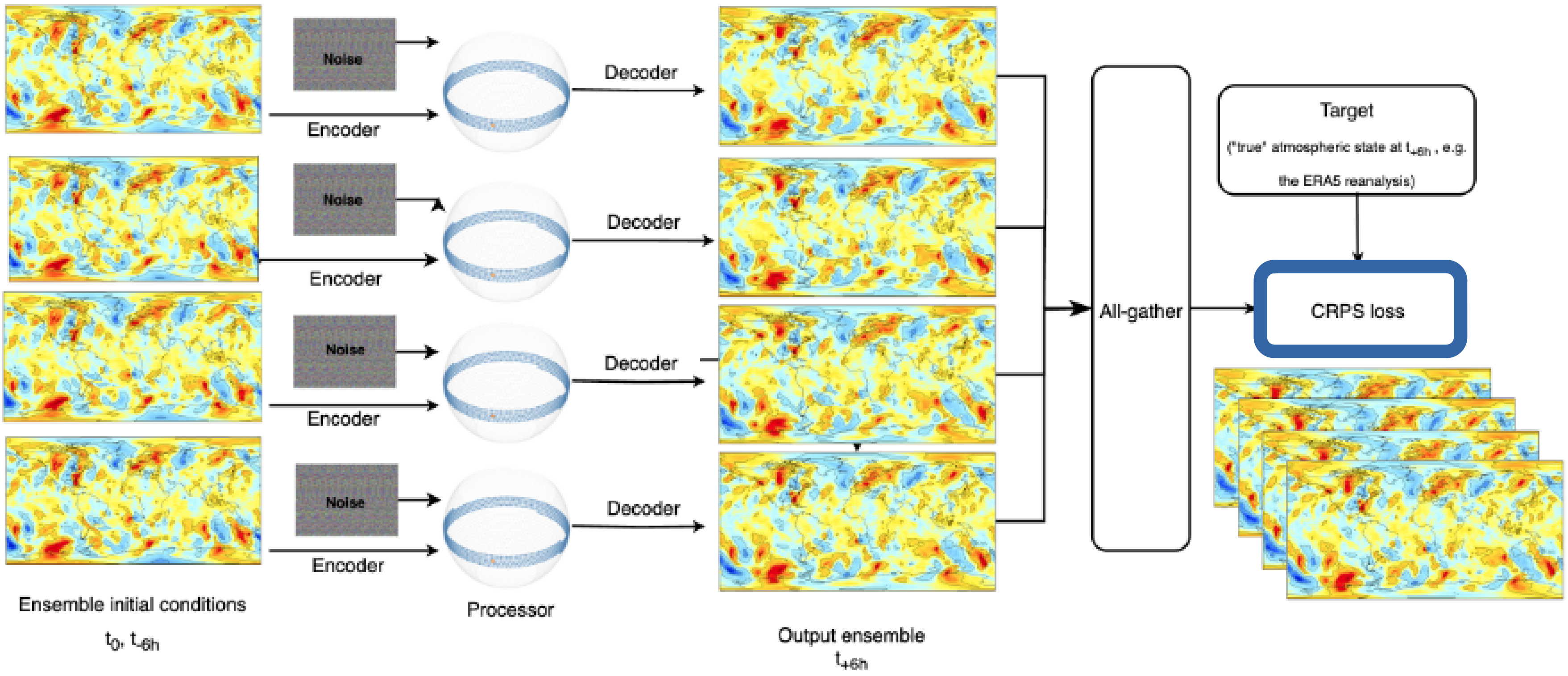
- **Model (epistemic) uncertainty** : How to estimate ?
  - Multi-model approach : multiple training, sampling models at different epochs
  - Monte carlo dropout
  - Bayesian neural networks



From Bano Medina et al., 2025

Score-based training

- Examples : AIFS-CRPS (Lang et al., 2024), Alet et al. 2025, Bonev et al., 2025, Kochkov et al., 2024, among others



From Lang et al., 2024

AIFS ENSEMBLE OPERATIONAL

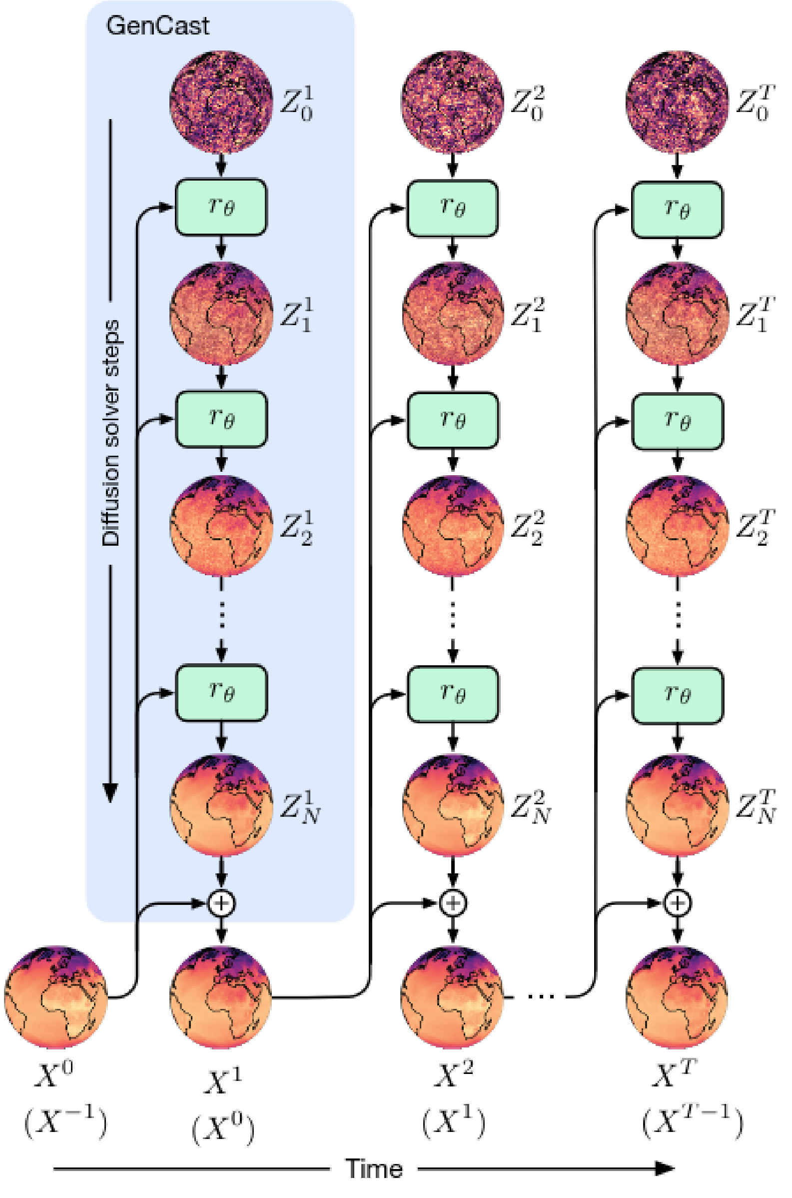
Generative models

From Price et al., 2024

- ➢ Diffusion (e.g., GenCast)
- ➢ Flow matching (Couairon et al., 2024)
- ➢ VAE (Oskarsson et al., 2024)

- ➢ Which type of uncertainty is represented ?
  - A change of paradigm ?

'Initial condition perturbation is not needed to generate weather trajectories with the correct dispersion.
Even if there is some uncertainty in the true weather state that affects the weather trajectories, the learned transition distribution should take this uncertainty into account, unlike traditional physical models' (Couairon et al., 2024)



Open questions

- Which approach is the best ?
- Need for a proper benchmark
- How do ML uncertainty compare with NWP uncertainty ?
- Most of the proposed solutions have been applied to global prediction at ~ 30 km resolution, **how do they transfer and perform at higher resolution ?**

→ Active research topic at the European scale (Machine Learning Pilot Project)
→ PhD Adrien Audren on Arome-AI EPS (Cifre Météo-France/Eviden)

Conclusions and future works

- Representing uncertainties is of utmost importance in weather forecasting
  - As models improve uncertainties remain
  - Ensemble forecasts are crucial for operational activities
  - Main limitations to UQ comes from : lack of understanding of all sources of uncertainty (especially model uncertainty) and constrained computational resources

- ML can offer new perspectives for representing uncertainties in physical models, by reducing some of the current limitations

- The rise of data-driven models is another lever for improved UQ, but also comes with new questions and challenges regarding the characterization of errors, deterministic vs probabilistic training, etc.

- Ensemble ML systems are still in their infancy