Ensemble reuse:
Impact of soil moisture guided ensemble sub-selection on the forecast skill of air temperature

Atmosphere Model, Chemistry Climate, Earth System Prediction, Climate Variability & Change, and Whole Atmosphere Working Group Meeting 2025

Feb. 5th

Daisuke Tokuda & Paul Dirmeyer (George Mason Univ.)

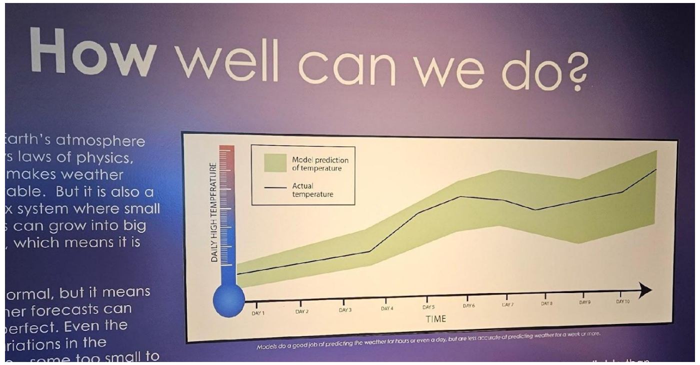
Introduction: Atmospheric chaos ⟶ S2S forecast

Chaos of the Earth system (in particular, the atmosphere)

Sensitivity to initial conditions

Ensemble forecasting

Multiple simulations with slightly different initial conditions around the best estimate



(1st floor of this building)

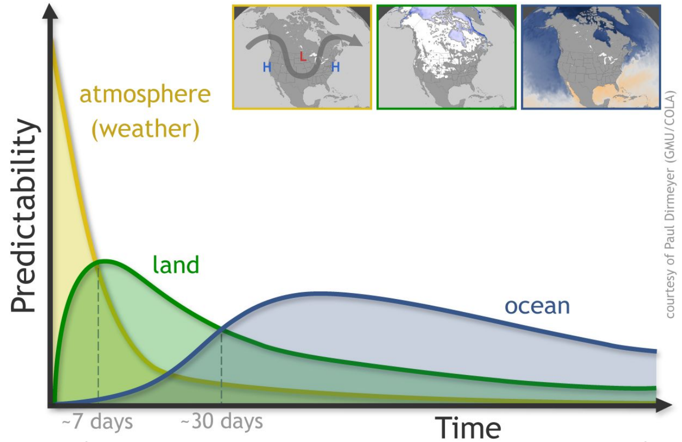
"Lifetime" of the initialization

Iterative initialization

For short-term forecast

Land initialization

For S2S-scale forecast

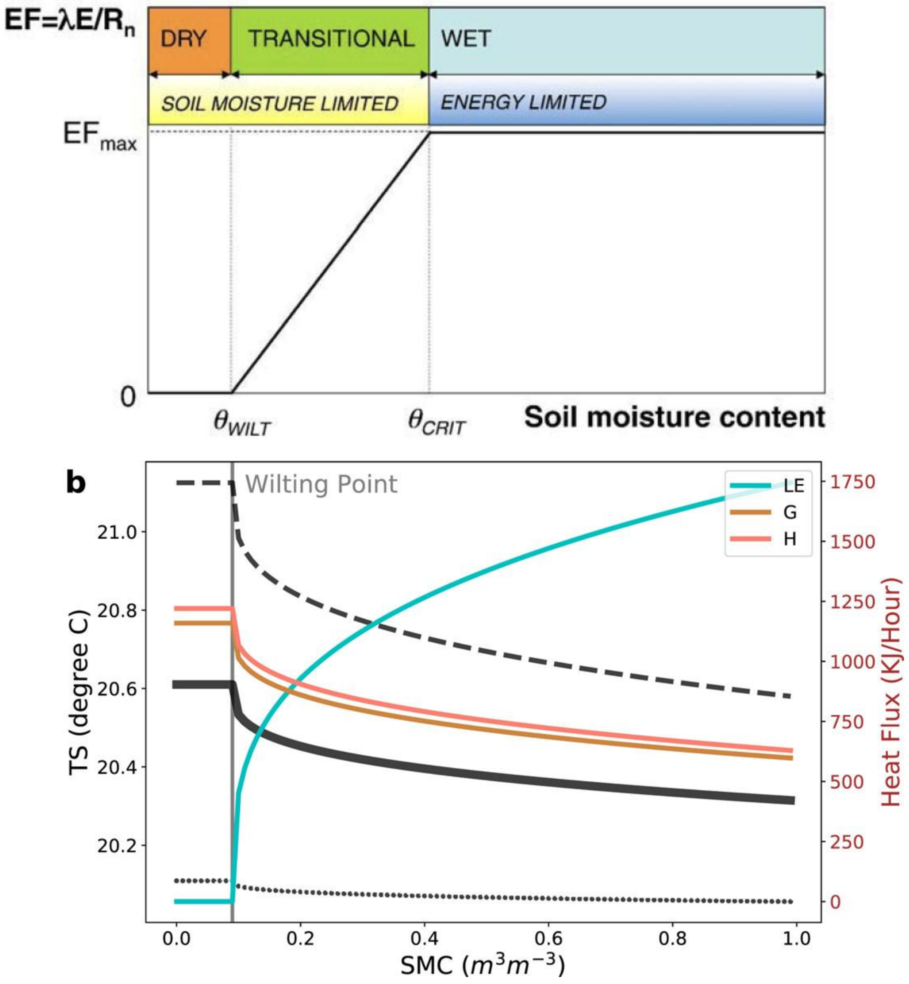


(Dirmeyer et al., 2015; NOAA)

Introduction: Soil moisture (SM) → Air temperature

Energy budget of land surface (skin):

Net radiation = Latent heat + Sensible heat + Ground heat



SM controls energy partitioning (Seneviratne et al., 2010)

- Soil moisture↓
- Latent heat↓ + Sensible heat↑
- Air temperature↑

In the context of heatwave, Dry SM

- Leads to extremely high temperature after large-scale atmospheric circulation (Miralles et al., 2014; Fernando et al., 2016)
- Affects subsequent predictability (Quesada et al., 2012)
- Induces a rapid temperature increase near the wilting point (Dirmeyer et al., 2021; Hsu et al., 2024)

The Earth's state changes constantly — *Fact*

The latest ensembles always outperform past ensembles — *Fact**?***

Ensembles initialized with the latest information

Ensembles initialized earlier but still continuing to forecast

Can we improve the latest forecasting skill with past ensembles?

This study's target: Daily maximum 2m air temperature in 1 to 4-week forecast

(Hereafter, *Air temperature*)

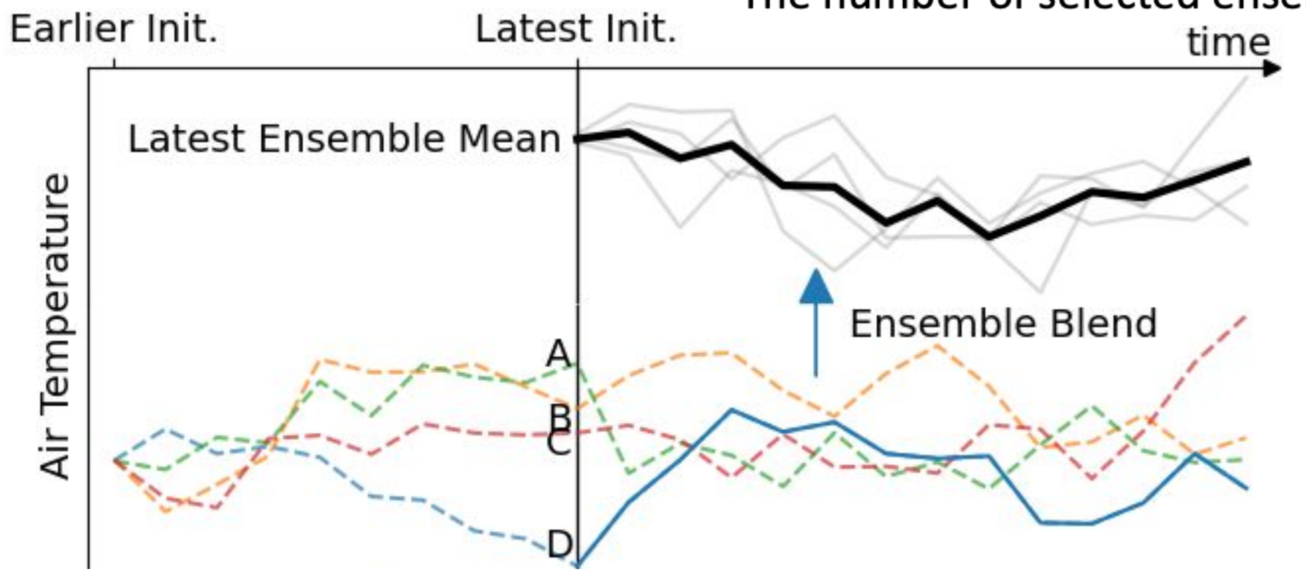
Proposed method: Ensemble reuse

At each grid cell,

2015-07-08 2015-07-15

The number of ensembles $N = 4$

The number of selected ensembles $M = 1$



SM comparison & ensemble sub-selection on 2015-07-15

2 parameters:

- How many ensembles to include
- How far back in time to select past ensembles

Optimized during the training period and tested during the testing period

Data

Reference (truth) data: NLDAS-2

- Temporal coverage: Summer season (June-July-August)
- Spatial coverage: 25°–53°N, 67°–125°W
  - Interpolated to each model's native grid
- Soil moisture: Top 28 cm

Forecast model: CESM2 (Richter et al., 2024) + 3 models in S2S database (Vitart et al., 2017)

- Soil moisture: Top 20 cm

Model	Ensemble #	Init. Interval	Forecast length	Grid #	Period	Train + Test #
CESM2	11	1 / week	45	29 x 47	2002 – 2022	17 + 4
ECMWF	11	2 / week	45	19 x 39	2004 – 2022	15 + 4
HMCR	11	1 / week	45	19 x 39	1991 – 2015	20 + 5
Météo-France	10	1 / week	46	19 x 39	1993 – 2017	20 + 5

- All of the variables are converted to percentile (%ile)
  - Forecasts are converted for each daily lead step to remove a drift

Methods

Parameter optimization: Brute-force search for each month

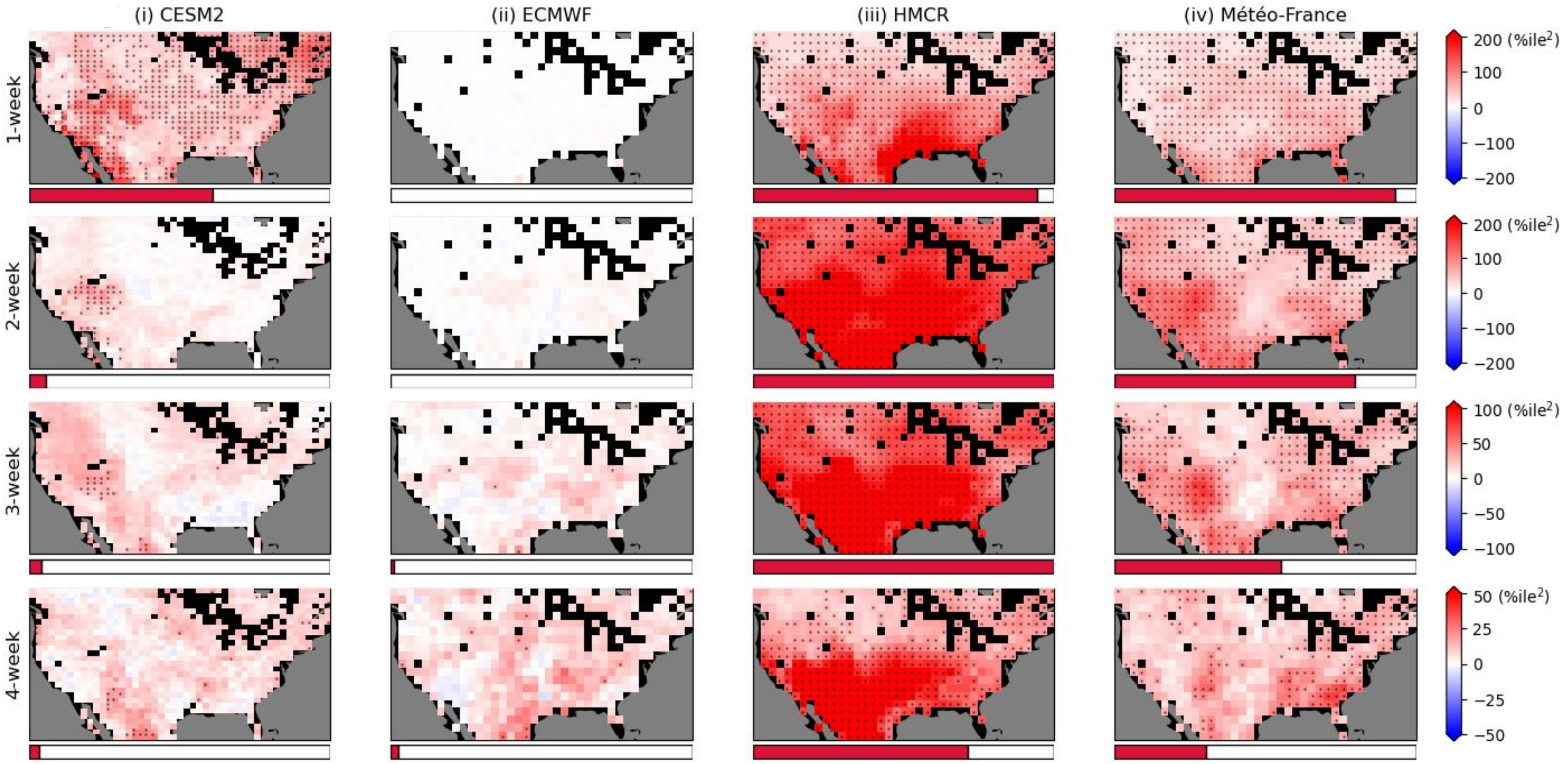
- Error metric for all parameter sets over the entire period
    - Mean Squared Error (MSE, unit: %ile^2)
    - Weekly scale: 0-7, 7-14, 14-21, 21-28 days
- Parameters are selected for each sample, each grid, each lead time, and each month

Uncertainty estimation: Bootstrapping method

- random split into training/testing years 100 times
- Results show
    - Mean of the 100 samples
    - Statistically significant if the 5th and 95th percentiles have the same sign

Results: Skill gain

ΔMSE (%ile^2) = Latest ensemble mean (LEM) − This study



Red (+): Improvement

▪: Statistical significance

Significant Improvement

Insignificant change

Significant Degradation

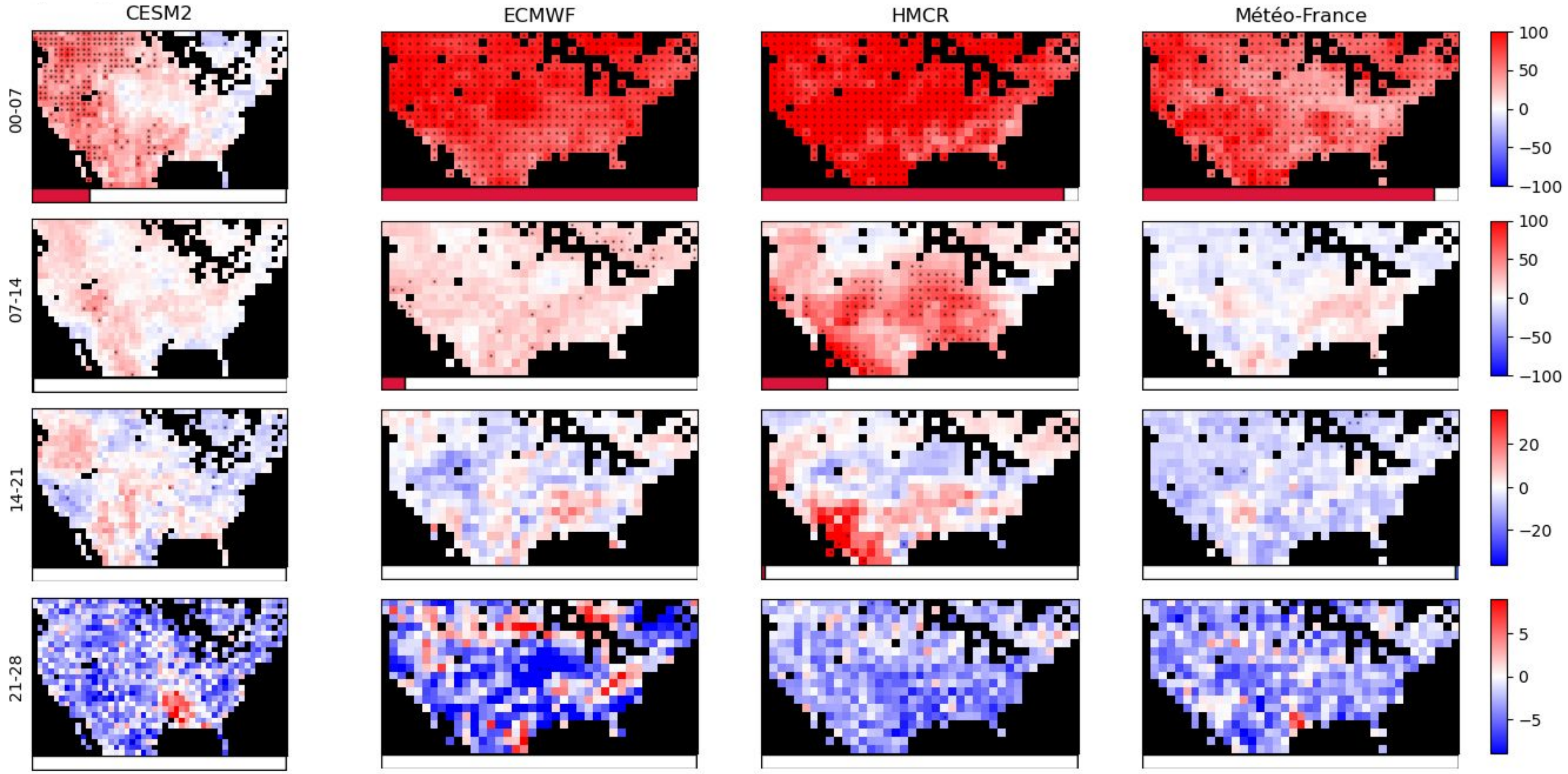
- Many grids are improved for the 1-week forecast
- The affected area has a spatial pattern and is reduced as the lead time gets longer
- ECMWF is insensitive to the proposed method

Results: Skill gain

It's natural that the realization error is improved if we increase the No. of ensembles.

Answer #1: Comparison b/w this study and reusing all ensembles of 1 week earlier

ΔMSE (%ile2) = 2-week ensemble mean ($2N$ ensembles) – This study ($N + M$)



+: Improvement

▪: 95% Statistical significance

- Selected ensembles ($N + M$) outperform the $2N$ ensembles
- Degradation is not statistically significant at almost all grids

Results: Skill gain

It's natural that the accuracy is improved if we increase the No. of ensembles.

Answer #2: MSE decomposition

$$Bias = \frac{1}{n}\sum e$$

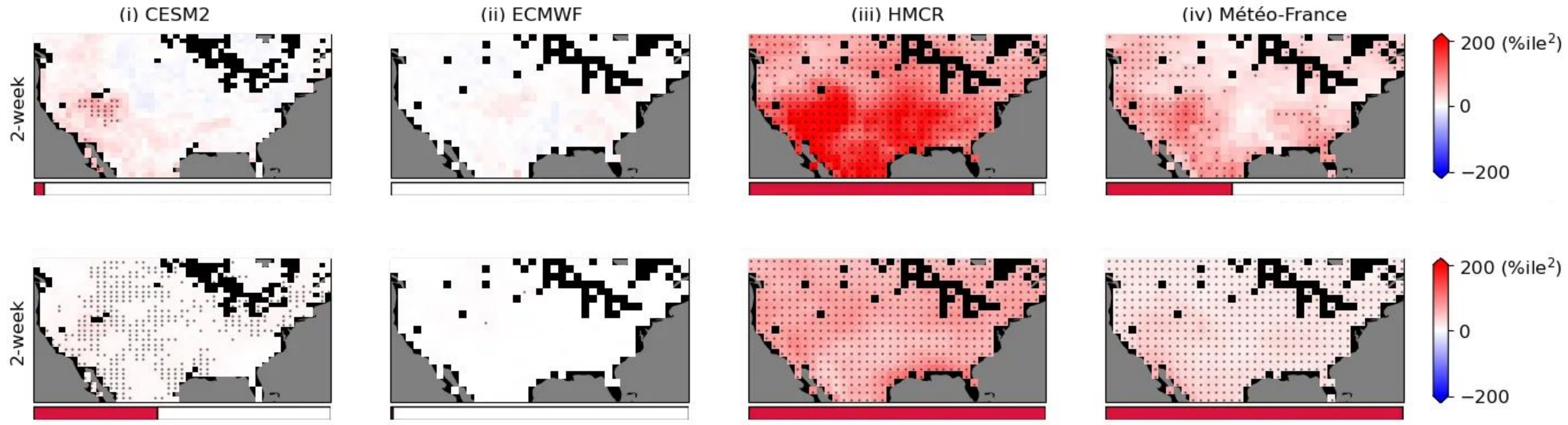
$$MSE = \frac{1}{n}\sum e^2 = \frac{1}{n}\sum\left((e - Bias) + Bias\right)^2 = \boxed{Bias^2} + \boxed{\frac{1}{n}\sum (e - Bias)^2}$$

Mean Error Realization Error (Variance)

y: Reference
$\hat{y}$: Forecast
$e = y - \hat{y}$
Use $\sum(e - Bias) = 0$

Mean Error

Realization Error

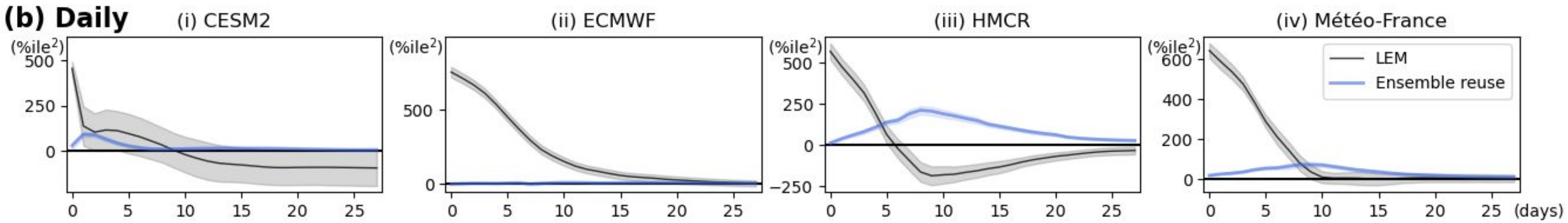


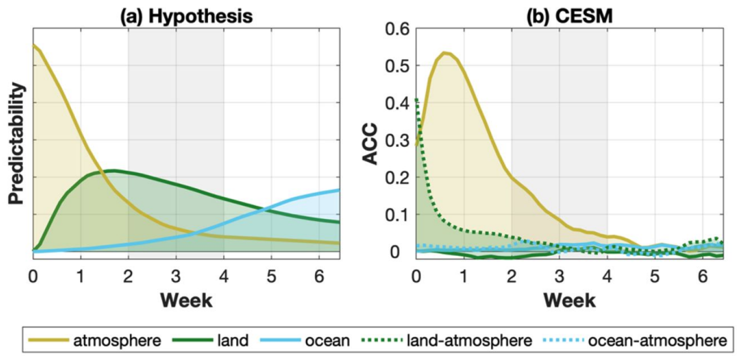
- (Only 3-week result shown, but the other weeks are similar)
- Realization error is decreased for many grids but its magnitude is minor
- Selective ensemble reuse also improves the mean error

Results: Hypothetic mechanism of ensemble reuse

Daily change in the skill gain averaged in the target region (%ile^2)

- Black: Reference climatology (no model) -> Latest ensemble mean
- Blue: Latest ensemble mean -> This study
  - +: Improvement
  - Shade: 1 STD among the 100 bootstrapping samples





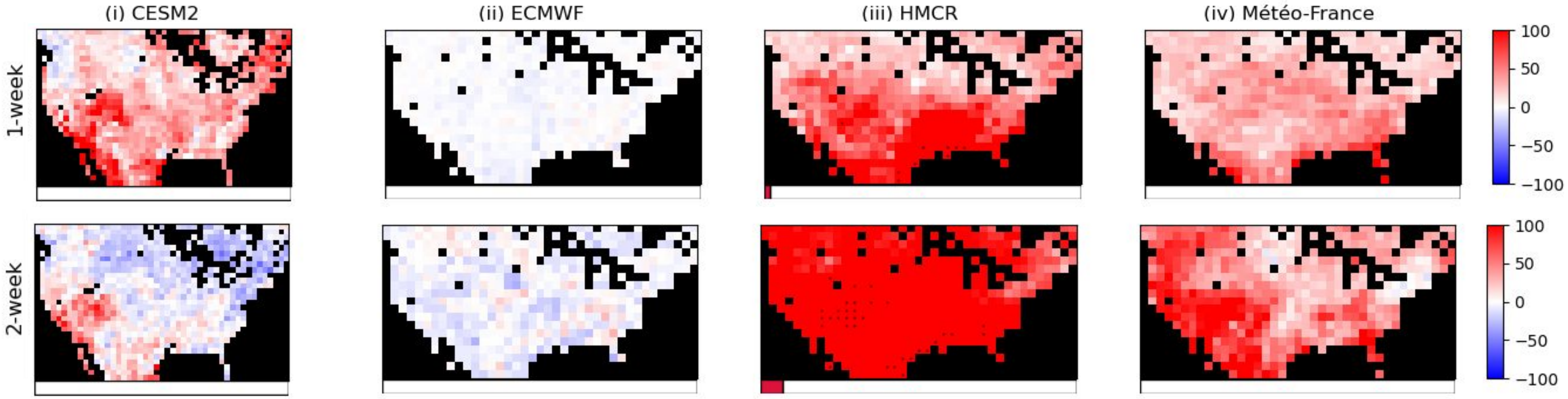
(Richter et al., 2024)

The similarity implies that the method supports (emulates?) land initialization

- The impact is affected by the current initialization
- Could be explained by the combination of SM-LE coupling and SM memory (Guo et al., 2011)

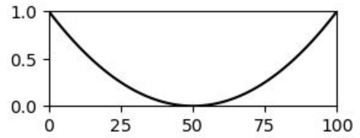
Results: Applicability to extreme cases

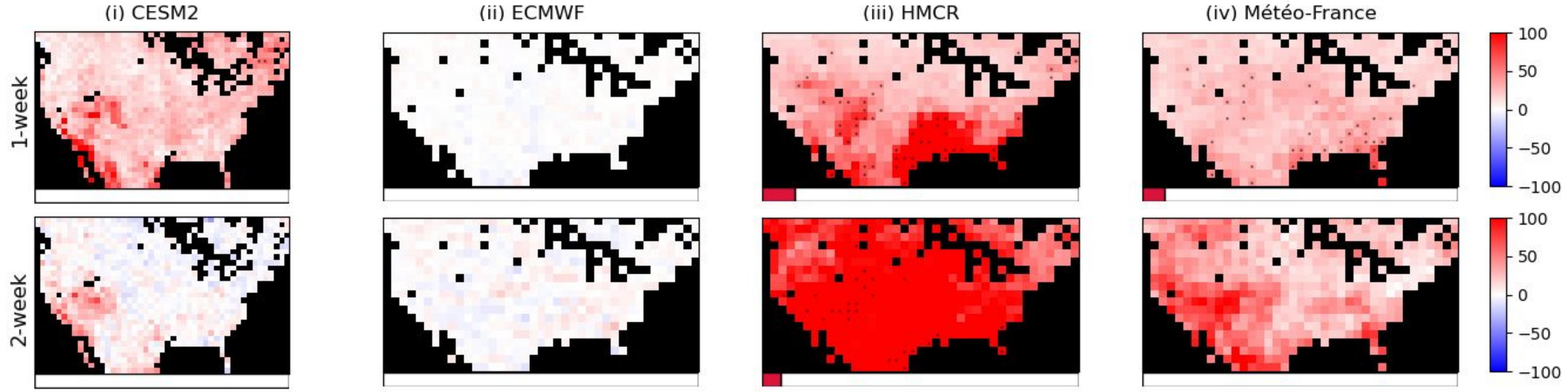
Skill gain for the 75-100 %ile of true air temperature (hotter side)



Loss function (MSE) weighted by true air temperature

$$\frac{1}{50^2}(Tair-50)^2$$





- 1-week forecast is improved, but the affected area is limited
- Flexibility and potential for fine-tuning are shown

Discussion: Comparison with other postprocessing (skipped)

Ensemble reuse can be seen as
- Initialization (or data assimilation?)
- Postprocessing

Note: Most recent comparative study uses only the latest forecast (Cho et al., 2022)

10 postprocessings in 4 classes (Yang et al., 2021)

To \ From	Deterministic	Probabilistic
Deterministic	• Regression • $\boxed{\text{Filtering}}$ • Resolution change	• Summarizing predictive distribution • Combining deterministic forecasts
Probabilistic	• $\boxed{\text{Analog ensemble}}$ • Method of dressing • Probabilistic regression	• Calibrating ensemble forecasts • $\boxed{\text{Combining probabilistic forecasts}}$

Ensemble reuse has advantages compared with the analog ensemble,
- Not require past forecasts of a “frozen” model (can be implemented in real-time)
- Robust against long-term change in climate and weather

Of course, generalization of the ensemble reuse with ML/AI would improve skill, which can be seen as data augmentation

Summary

Can we improve the latest forecasting skill with past ensembles?

This study's target: Daily maximum 2-m air temperature

—Yes, we can!

Ensemble reuse: Blend the latest ensembles with earlier ensembles selected with the soil moisture forecastability at the same date

- Improves the forecast skill for CESM2, HMCR, and Météo-France for 1 to 4-week
  - Affected area shrinks as the leadtime gets longer
- Does not affect much ECMWF
- Does not degrade the skill of the latest ensemble mean

- Could be explained by additional improvement in land initialization
  - In addition to skill improvement, the method would emulate the upper boundary of the initialization improvement under the current model
- Would be enhanced with ML/AI (⟶ can be used as Data augmentation)

Thank you for listening!