Center for Western Weather and Water Extremes
SCRIPPS INSTITUTION OF OCEANOGRAPHY AT UC SAN DIEGO

West-WRF Physics and AI: CW3E's Latest Developments for the Prediction of Atmospheric Rivers and Extreme Events

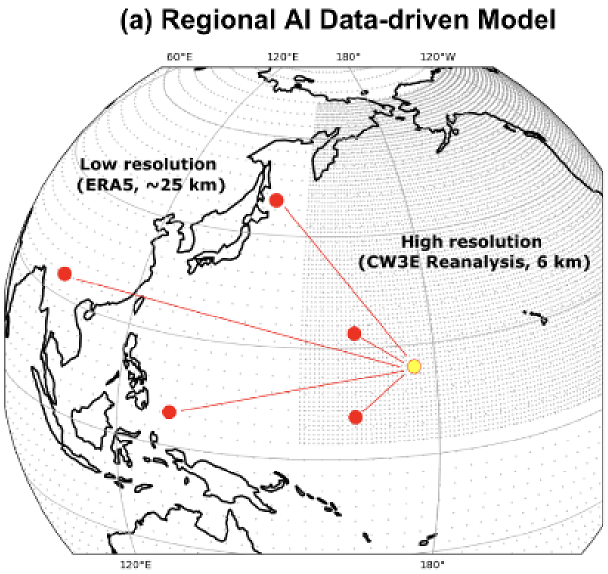
Luca Delle Monache
Director of Research, Center for Western Weather and Water Extremes (CW3E)
Scripps Institution of Oceanography (SIO)
University of California San Diego (UCSD), La Jolla, California, USA

2026 AR Recon workshop and 2nd Observational campaigns workshop for better weather forecasts
30 June 2026

Sponsor Acknowledgements
- AR Program (California DWR)
- FIRO Program (USACE)
- SAFARI (US ONR)

UC San Diego | SCRIPPS INSTITUTION OF OCEANOGRAPHY

WEST-WRF AI: A REGIONAL AI DATA-DRIVEN MODEL



Conceptual diagram outlining the development of the AI data-driven model based on Graph Neural Networks (GNNs)

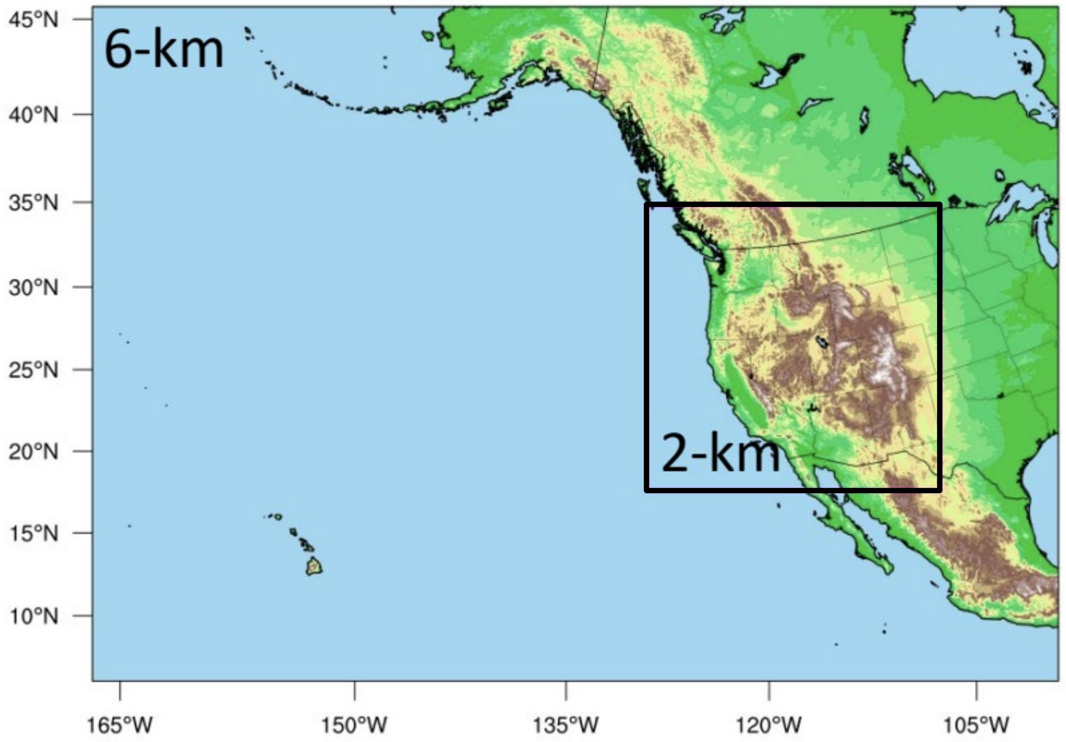
We are developing high-resolution, regional (and global) AI data-driven models to better predict and understand atmospheric rivers (ARs) and associated extremes. This project leverages:

- ECMWF's Artificial Intelligence Integrated Forecasting System (AIFS; Lang et al. 2024), global and regional modeling capabilities, via the *anemoi* framework
- Key collaborators: ECMWF (Matthew Chantry, Florian Pappenberger, Mario Santa-Cruz) and Norwegian Meteorological Institute (Thomas Nipen)
- CW3E's new 41-year regional reanalysis at 2- and 6-km resolution over the western US and Northeast Pacific, based on West-WRF physics-based model
- ECWMF's ERA5 reanalysis at 31-km resolution for the rest of the global domain

Baño-Medina et al., 2025: A regional high resolution AI weather model for the prediction of atmospheric rivers and extreme precipitation. *npj Climate and Atmospheric Science*
Rouzegari et al., 2026: West-WRF AI 2-km: high-resolution prediction of integrated vapor transport and precipitation. *npj Artificial Intelligence (AI for NWP), in preparation*

CW3E'S 43-YEAR HIGH-RESOLUTION REANALYSIS (WEST-WRF Physics)

- Dynamical downscaling of ECMWF's ERA5 (0.25º) with West-WRF
- West-WRF is a regional physics-based model tailored for the prediction of ARs and associated extremes
- Additional specialized inputs
  - ERA5 Land (initial soil and surface fields, 0.1º)
  - OSTIA SST and sea ice (0.1º)
- 100 vertical levels (12 below 1000 m)
- The 6-km domain run with grid nudging for wind speed, temperature, and moisture
- 43-years: 1982-2025

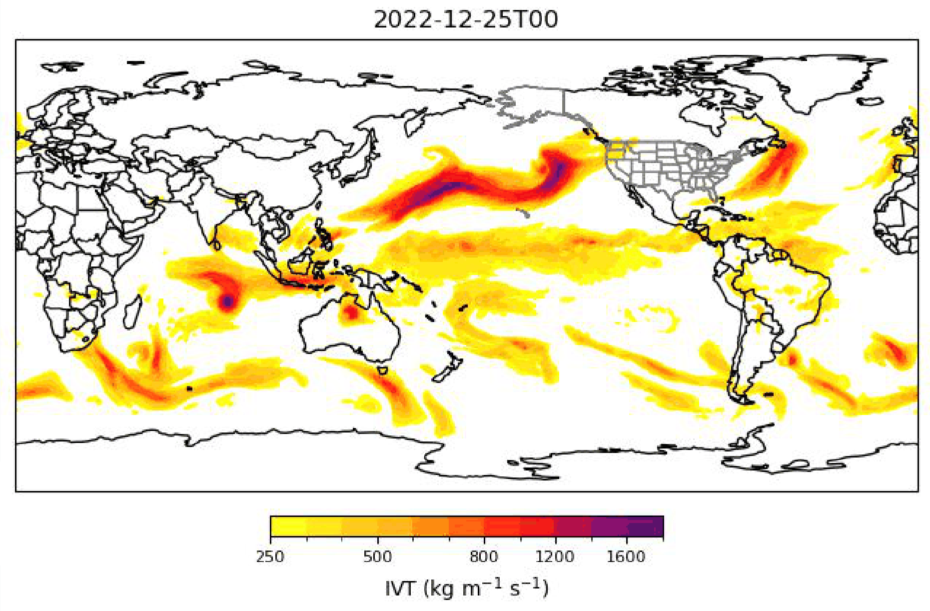


NOTE So far, we have used 8 years for training and 3 years for independent testing

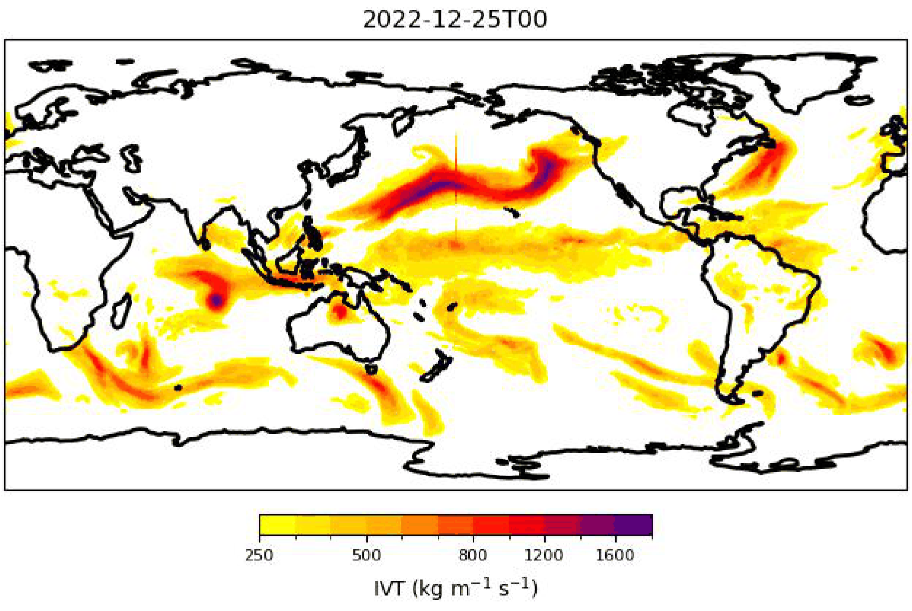
PI: Dan Cayan
Steinhoff et al. 2026: *in preparation*

NEW YEAR'S EVE ATMOSPHERIC RIVER OF 2022 (IVT)

ERA5 (~ 31-km)



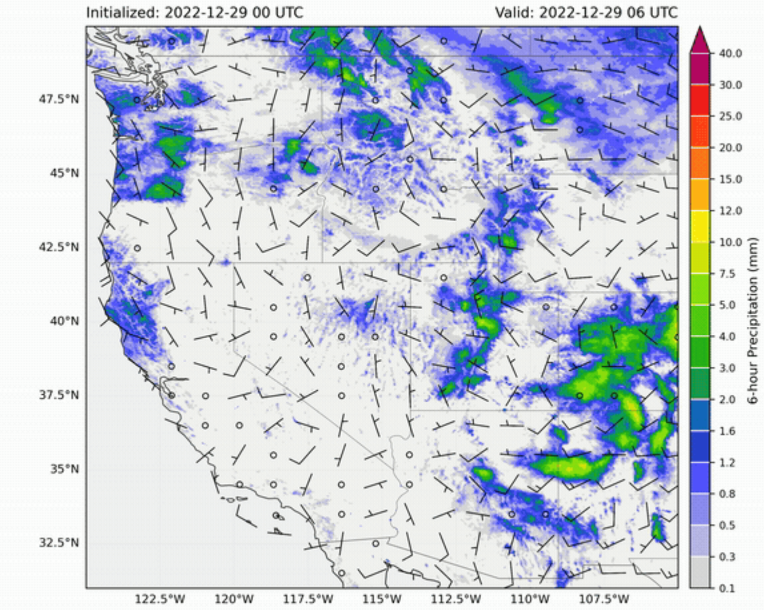
West-WRF AI (~31-km)



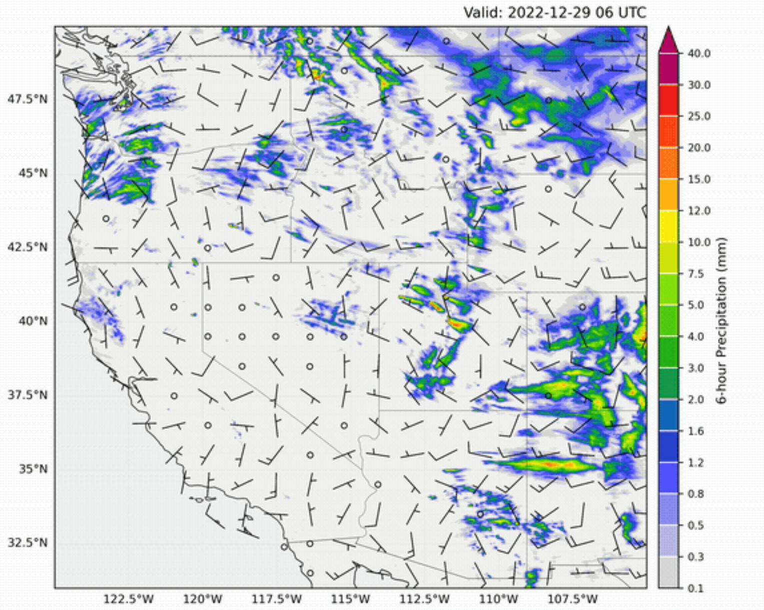
New Year's Eve Atmospheric River of 2022: Integrated Vapor Transport (IVT, kg m^{-1} s^{-1}), for a 7-day prediction with ERA5 (left) and the AI global model (right)

NEW YEAR’S EVE ATMOSPHERIC RIVER OF 2022 (PRECIPITTATION)

West-WRF AI (2-km)



West-WRF Reanalysis (2-km)

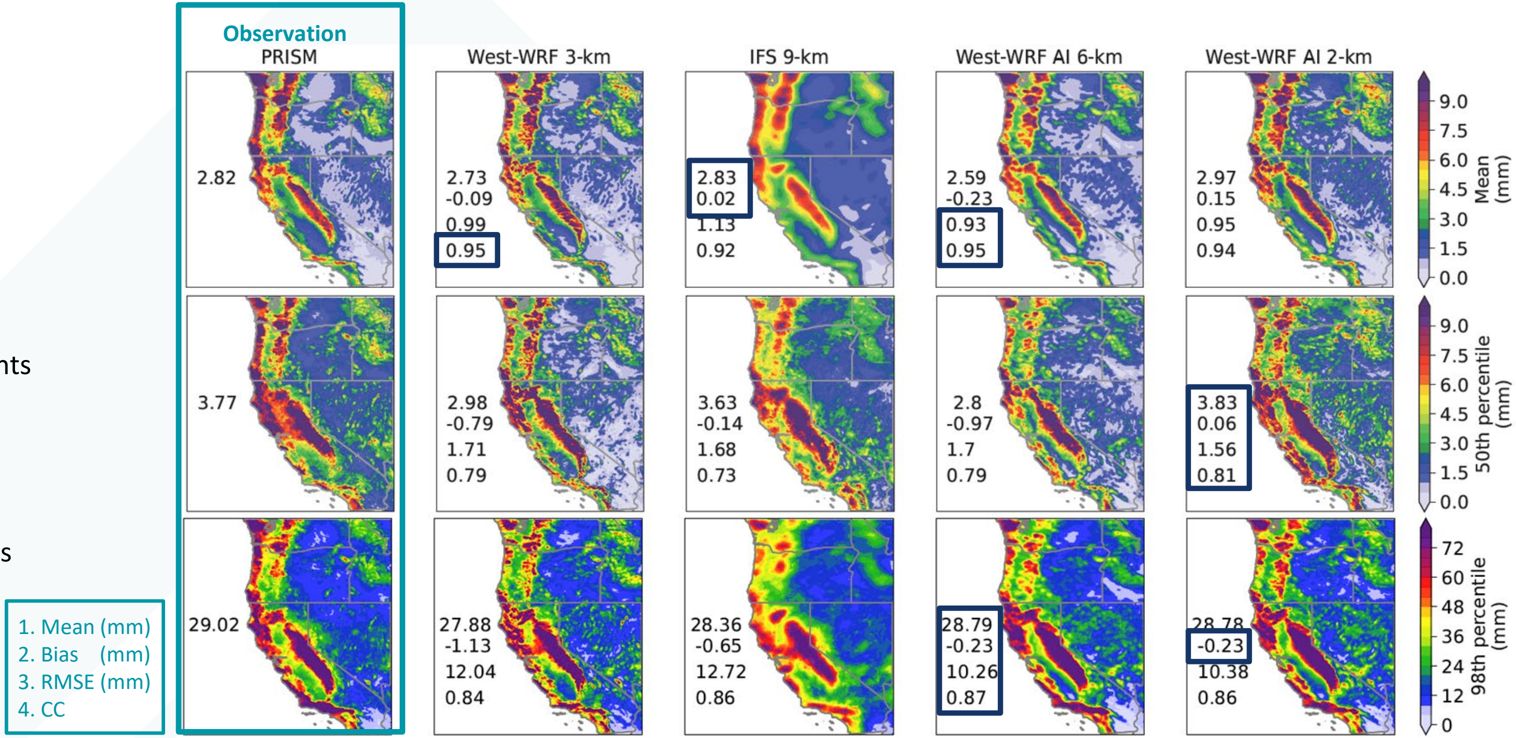


New Year’s Eve Atmospheric River of 2022: Integrated Vapor Transport (IVT, kg m^{-1} s^{-1}), for a 7-day prediction with ERA5 (left) and the AI global model (right)

PERFORMANCE OVER SPATIAL DOMAIN (PRECIPITATION, DAY 4)

- Test period: Winters 2020–2023
- Observation: PRISM 24-hour accumulated precipitation (4-km)
- Lead time: 4 days

- Mean precipitation
- Moderate precipitation events (50th percentile)
- Extreme precipitation events (98th percentile)

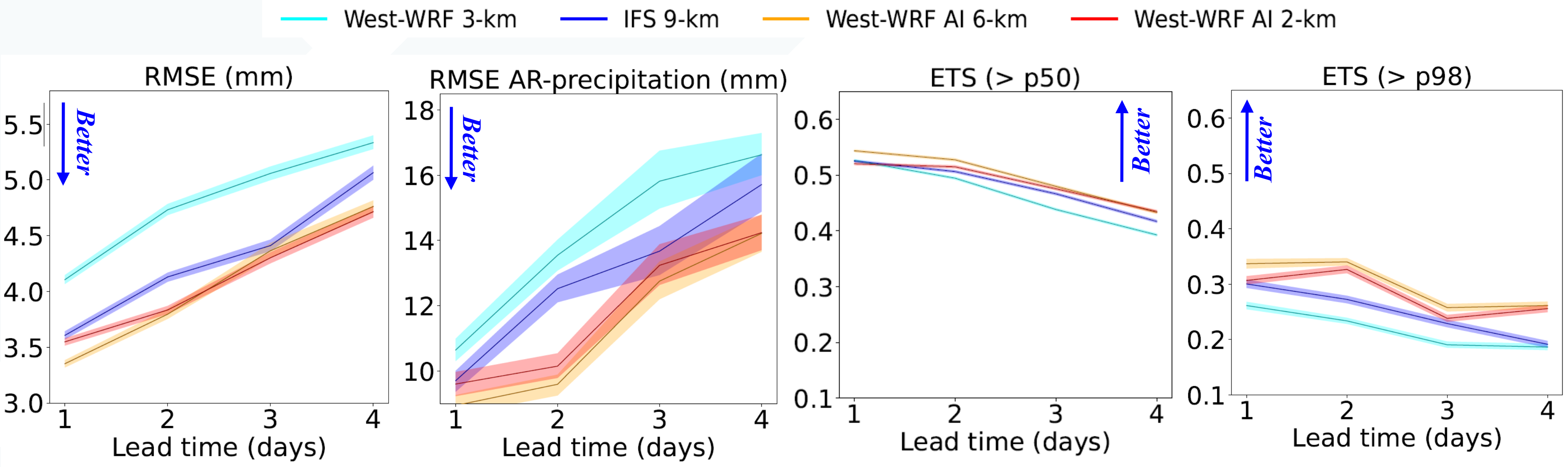


West-WRF AI (6- and 2-km): Improved representation of spatial structure and precipitation extremes

PERFORMANCE OVER LEAD TIME (PRECIPITATION)

- **Test Period:** Winters 2020–2023
- **Observation:** PRISM 24-hour accumulated precipitation (4-km)
- **Event Thresholds:** PRISM daily climatological percentiles (p50, p98)

**ETS: Equitable Threat Score*

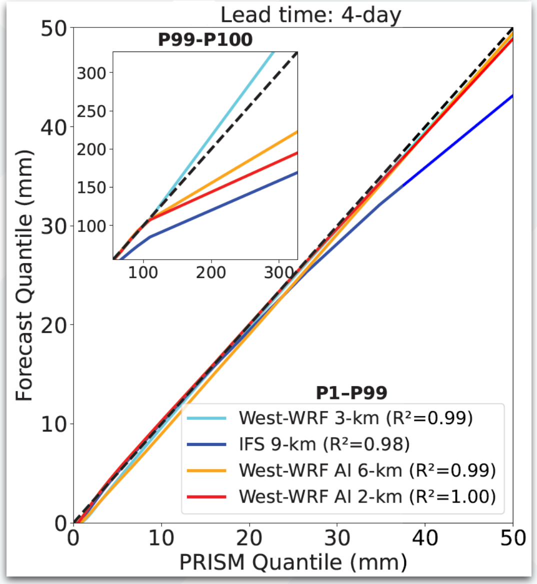


West-WRF AI 2-km: reduced forecast error and better/comparable detection of moderate and extreme events

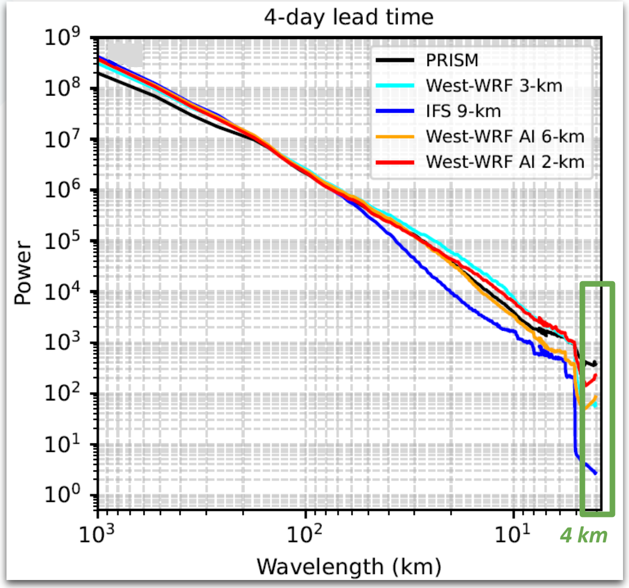
PERFORMANCE OVER PREDICTIVE DISTRIBUTION (PRECIPITATION)

- Test Period: Winters 2020–2023
- Observation: PRISM 24-hour accumulated precipitation (4-km)
- Lead time: 4 days

Quantile-Quantile (Q-Q) plot



Radially Averaged Power Spectral Density (RAPSD)

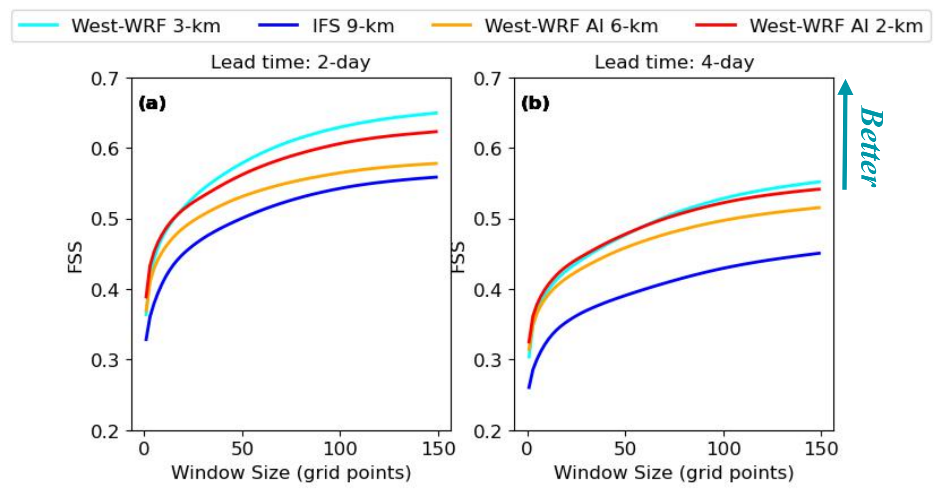


West-WRF AI (6- and 2-km): close observed distribution matching, physis version a bit better for p99-100, 2-km AI model better for power spectra representation at higher resolution

PERFORMANCE ACROSS SPATIAL SCALES AND INTENSITY (PRECIPITATION)

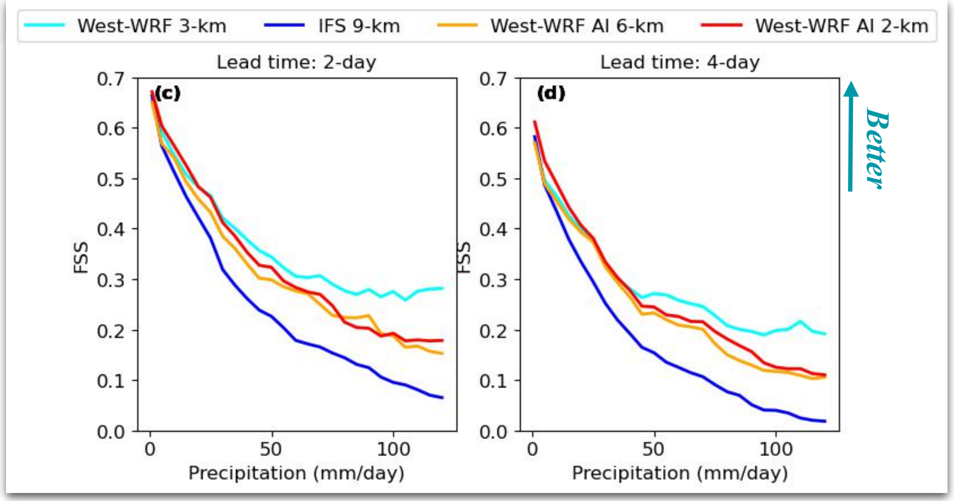
- Test Period: Winters 2020–2023
- Observation: PRISM 24-hour accumulated precipitation (4-km)
- Lead time: 2 days & 4 days

Fractional skill score (FSS)
Threshold: 20 mm/day



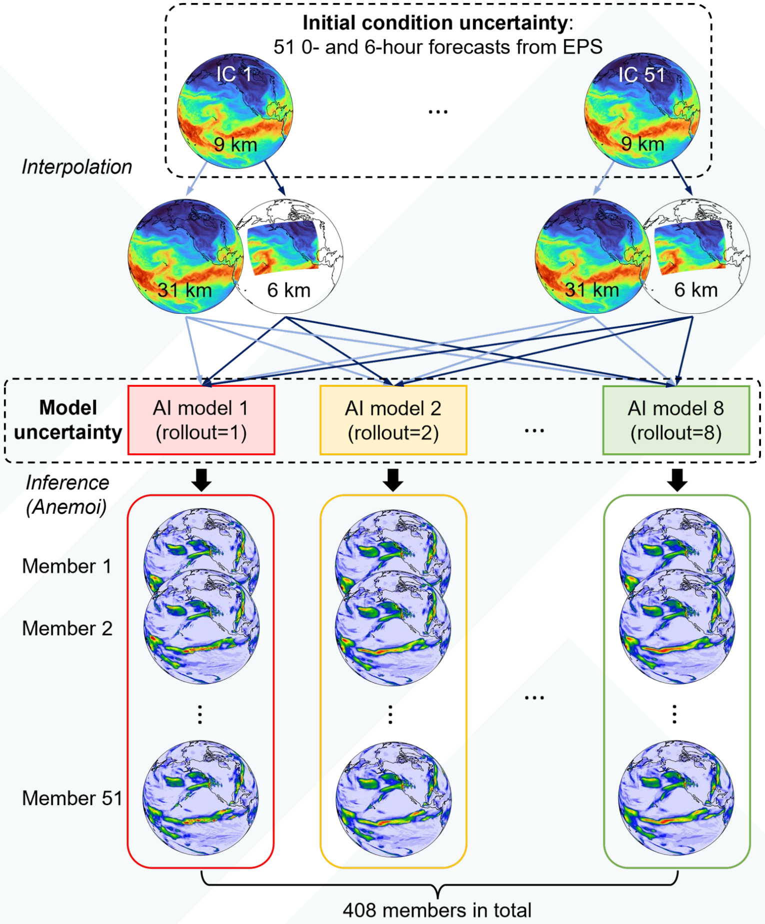
(grid × 4 = km) *(grid × 4 = km)*

Fractional skill score (FSS)
Window: 50 km



West-WRF AI 2- and 6-km comparable, physics version better for heavy precipitation

PROBABILISTIC FORECASTING: WEST-WRF AI ENSEMBLE DESIGN



INITIAL CONDITION UNCERTAINTY
(51 samples)

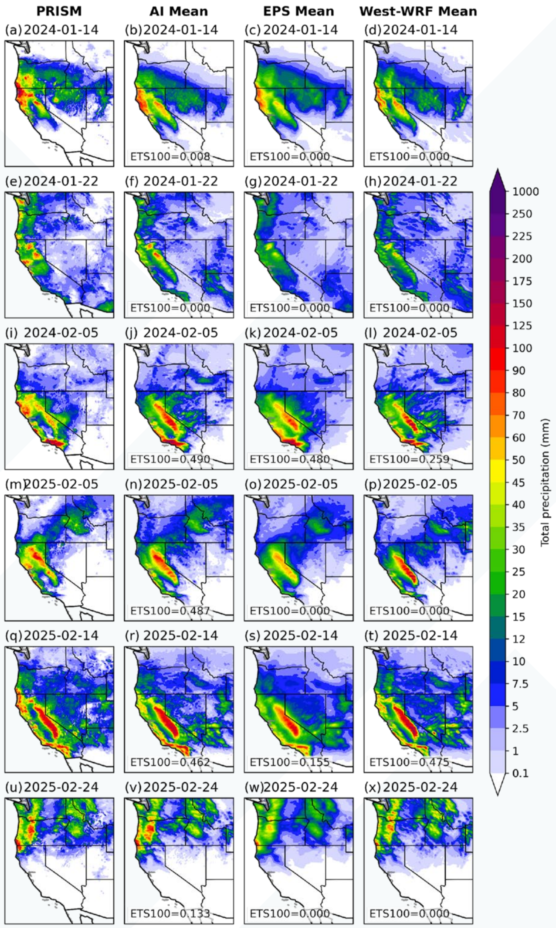
MODEL UNCERTAINTY
(8 samples)

ENSEMBLE
(51 x 8 = 408 ensemble members)

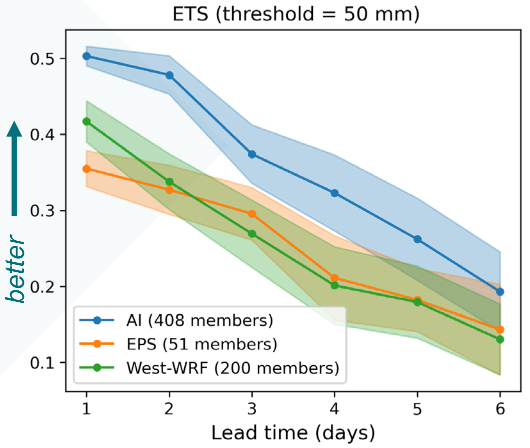
Nie et al., 2026: High-resolution ensemble forecasting of regional extreme precipitation via hybrid physics-AI modeling. *npj Climate and Atmospheric Science, submitted*

CASE STUDIES (WEST-WRF AI ENSEMBLE, 408 members, 6-km)

Lead Time:
Day 3

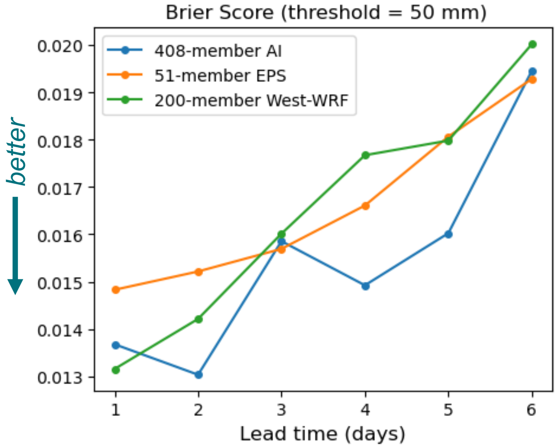


Equitable threat score (ETS)



Shading: average ETS across the 408 members
Solid line: std. dev. of ETS across the 408 members

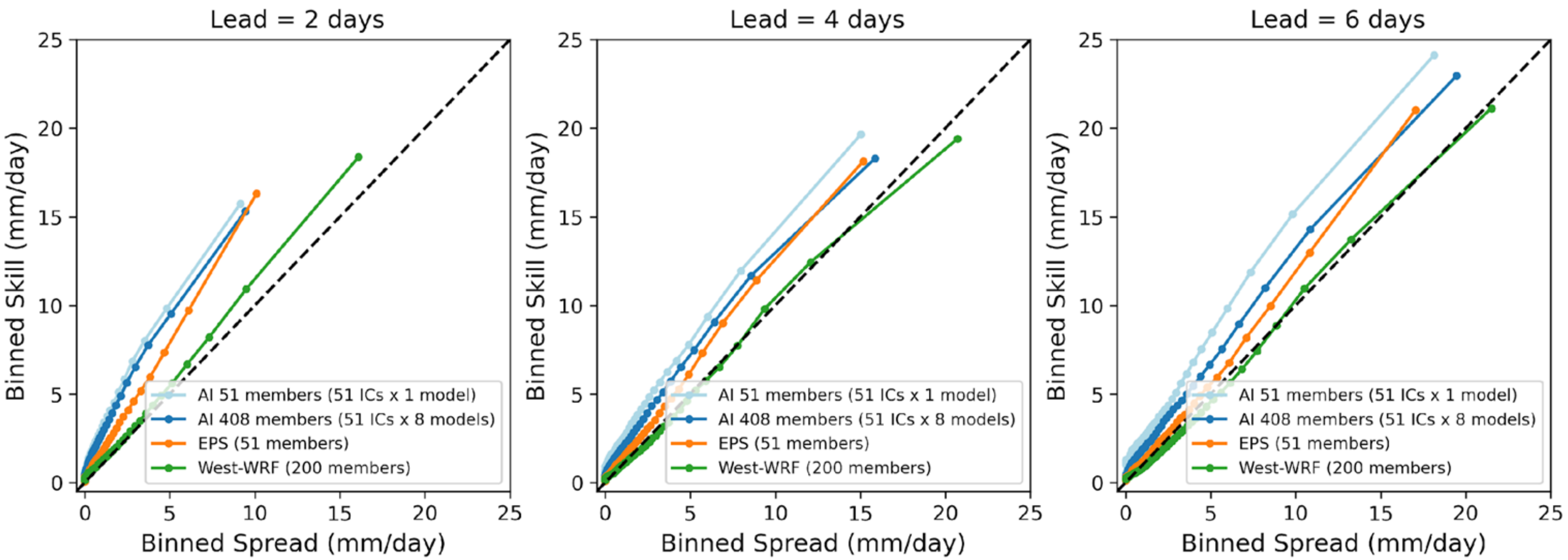
Brier Score



The AI-based ensemble outperforms physics-based ensembles in predicting these six extreme precipitation events

Nie et al., 2026: High-resolution ensemble forecasting of regional extreme precipitation via hybrid physics-AI modeling. *npj Climate and Atmospheric Science, submitted*

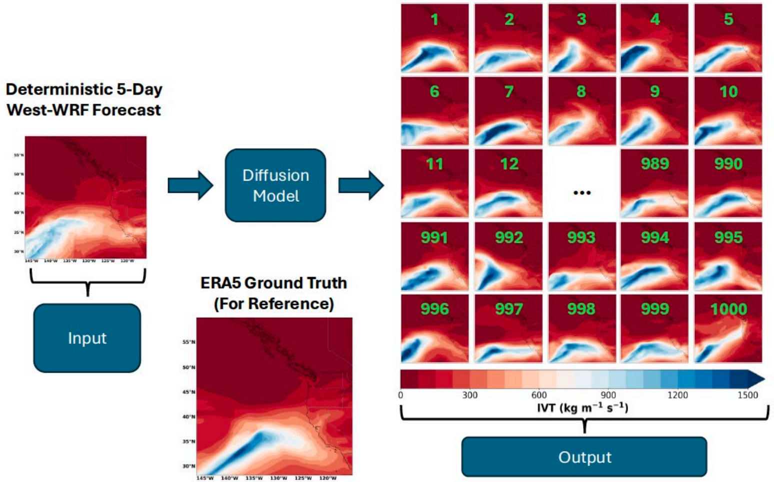
PRELIMINARY ASSESSEMENT OF AI ENSEMBLE

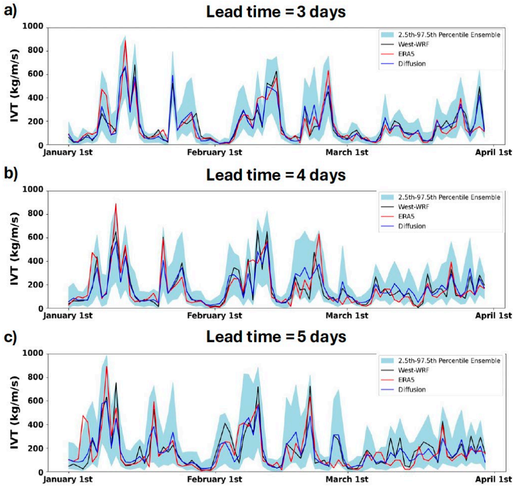


- Metric: binned-spread/skill plot for 24-h accumulated precipitation over the Western US
- Testing period: December-January-February 2023–2024 and 2024–2025
- The AI ensembles are underdispersive; West-WRF has the best spread/skill relationship

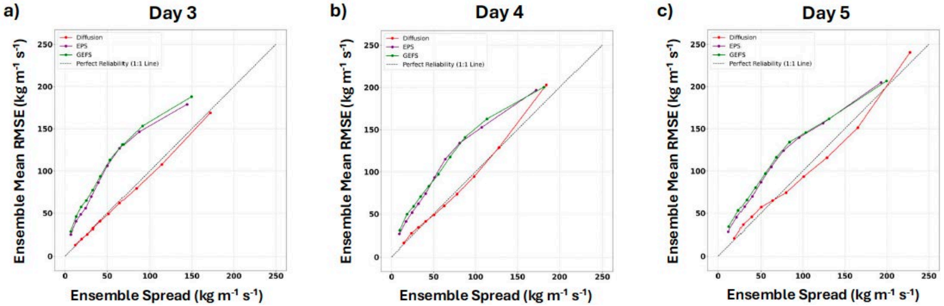
Nie et al., 2026: High-resolution ensemble forecasting of regional extreme precipitation via hybrid physics-AI modeling. *npj Climate and Atmospheric Science, to be submitted*

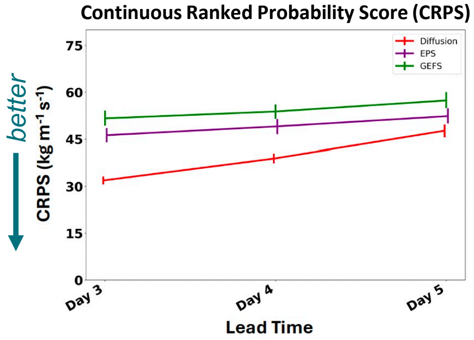
A DIFFUSION MODEL FOR THE PREDICTION OF IVT





IVT predictions at San Francisco, CA from January 1st , 2017 to March 31st , 2017





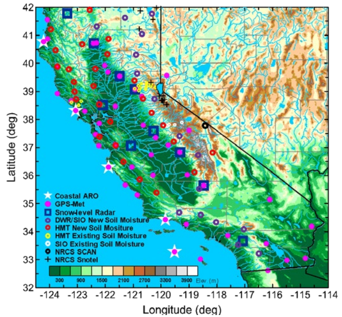
Higgins et al., 2026: Generating a 1000-member ensemble of integrated water vapor transport forecasts with diffusion. *AI for the Earth Systems, conditionally accepted*

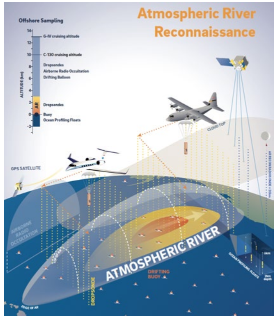
THANK YOU!

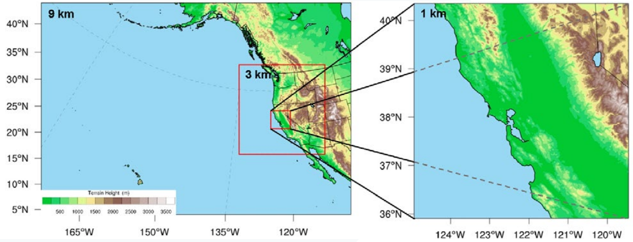
CW3E STRATEGIES TO IMPROVE AR-RELATED PRECIPITATION FORECASTS

(CW3E 2024-2029 Strategic Plan)

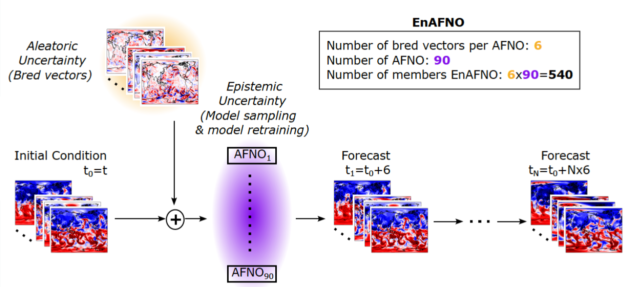
Observations and Data Assimilation To better understand key physical processes, to improve forecasts initial conditions, and to verify the skill of predictive capabilities



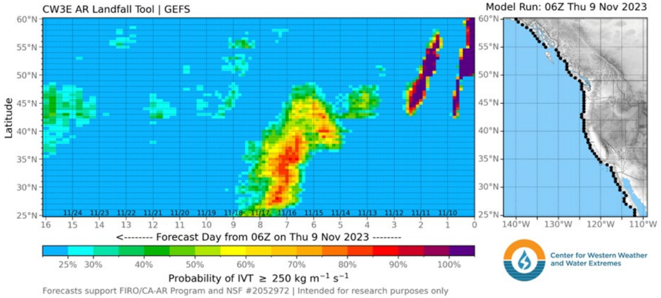




AR and Precipitation Science and Physics-based Modeling Dynamical models based on the state-of-the-science knowledge of physics and numerical procedures to predict ARs and associated precipitation



Machine Learning to leverage physics-based models and further improve their skill with a focus on ARs and extreme events



Decision Support Tools To synthetize vast amount of data from a range of platforms and methods, to support effective decision-making of end users

Center for Western Weather and Water Extremes

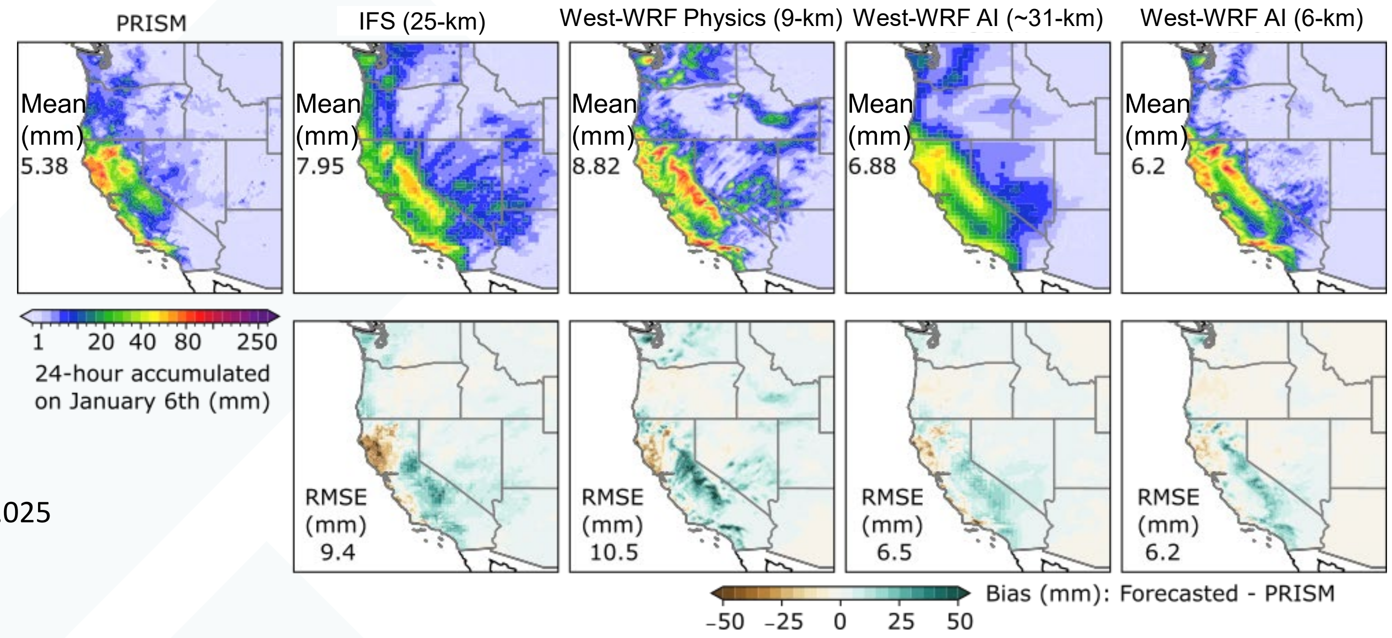
WEST-WRF AI @ 6-KM: CASE STUDY, AR3 ON 6 JANUARY 2023
24-h ACCUMULATED PRECIPITATION, DAY-3 FORECASTS

Model Verification
- PRISM (ground-truth) resolution: 4 km
- 24-hour accumulated precipitation

Model Hyperparameters

B1
# of Epoch: 34
Learning rate: 4e-6
Rollout: 1
Total precipitation loss weight: 0.025
IVT loss weight: 0
Regional domain weight: 0.25
Regional refinement: 9
Global Refinement: 7

B2
# of Epoch: 8
Learning rate: 3e-6
Rollout: 8



WEST-WRF AI @ 2-KM: CASE STUDY, AR3 Jan 11-14, 2023
3-DAY ACCUMULATED PRECIPITATION, 3-DAY FORECASTS

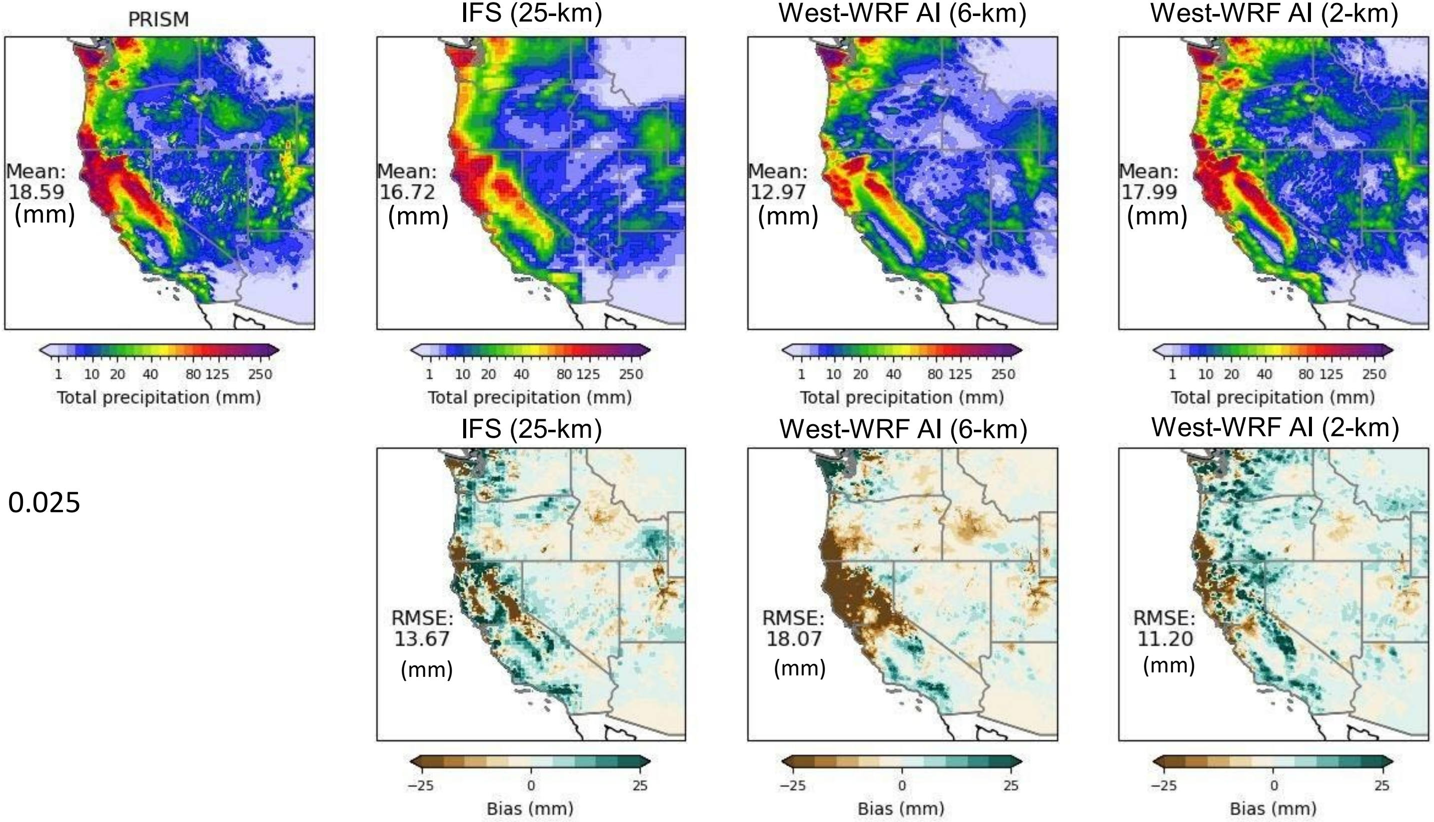
Model Verification
- PRISM (ground-truth) resolution: 4 km
- 24-hour accumulated precipitation

Model Hyperparameters

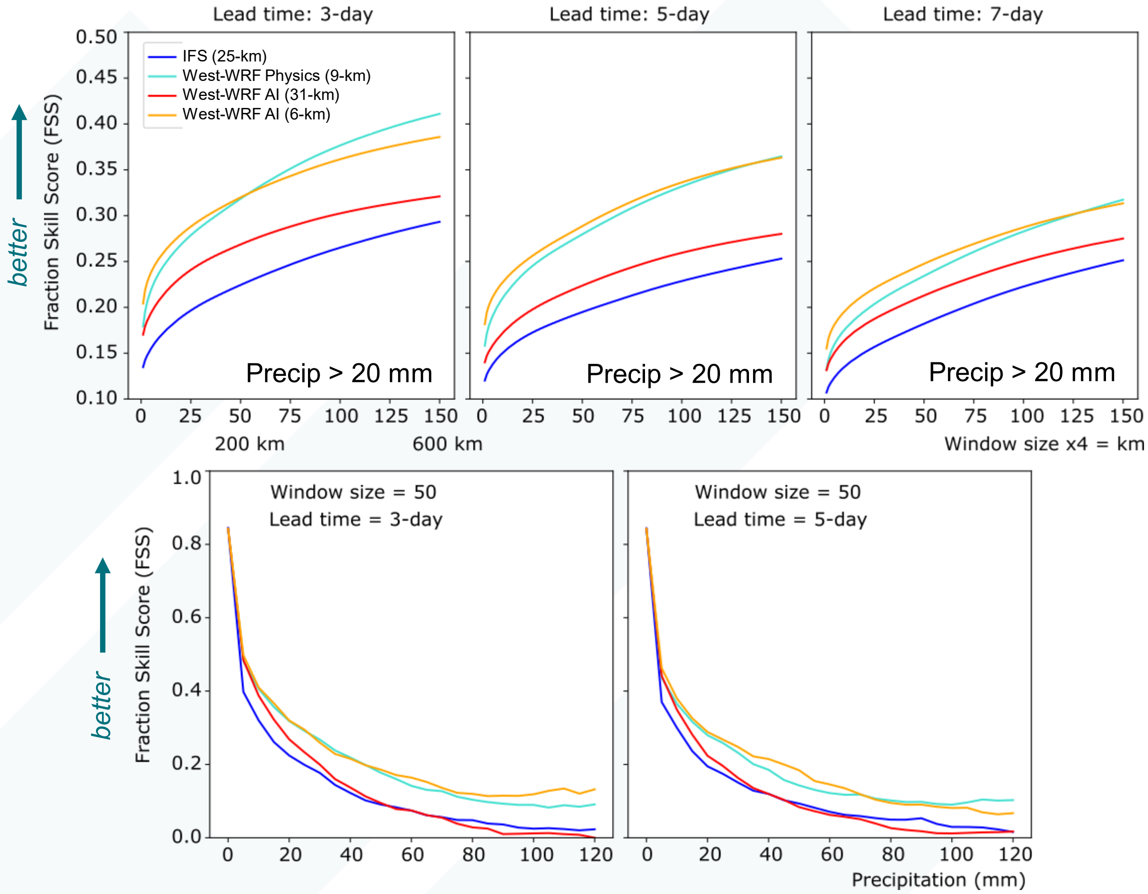
<u>B1</u>
# of Epoch: 38
Learning rate: 4e-6
Rollout: 8
Total precipitation loss weight: 0.025
IVT loss weight: 0
Regional domain weight: 0.25
Regional refinement: 9
Global Refinement: 7

<u>B2</u>
# of Epoch: 9
Learning rate: 8e-7
Rollout: 8

Case study: AR3 (January 11-14, 2023)



WEST-WRF AI @ 6-KM: VERIFICATION METRICS (NEIGHBORHOOD VERIFICATION)

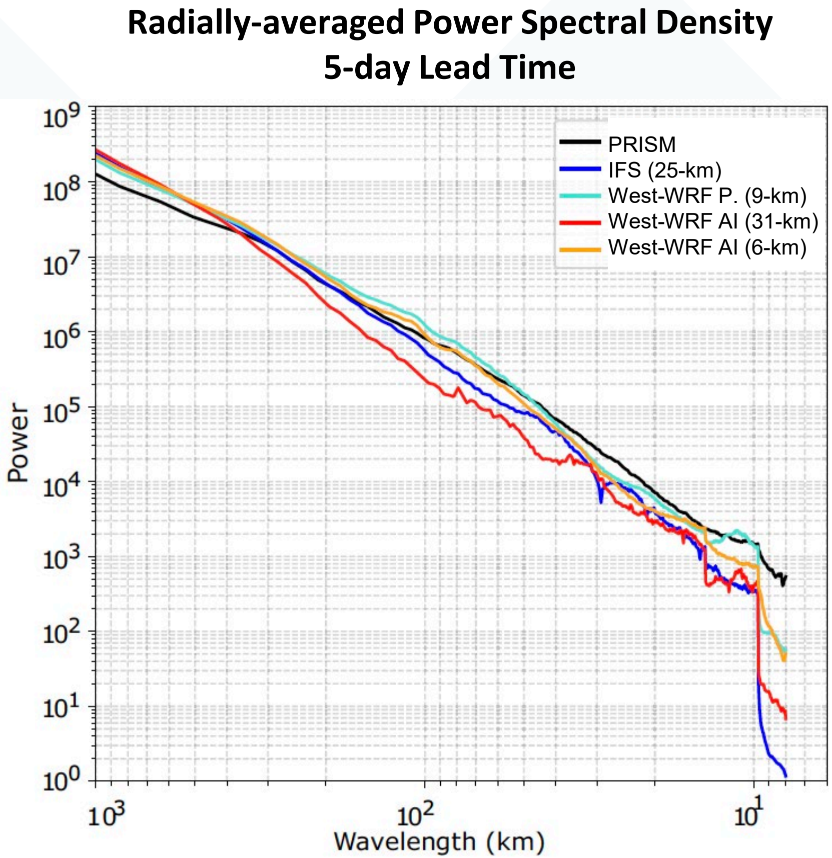


Training Period
Nov 1 through Mar 31
1979-2020 (ERA5)
2012-2020 (CW3E 41-year reanalysis)

Testing period
(data not seen during training)
Nov 1 through Mar 31
2020–2021, 2021–2022, and 2022–2023

Bano-Medina et al. (2025)

WEST-WRF AI @ 6-KM: WHAT SCALES ARE THE AI MODELS RESOLVING?



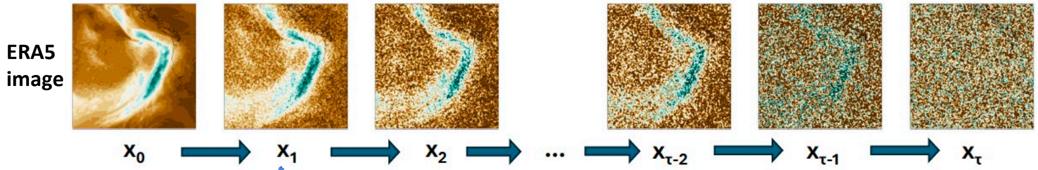
- Smoothness and underestimation of precipitation for both the IFS and AI 31-km models, with low values at higher frequencies
- The regional models closely follow the power spectral density of the PRISM observational target, showing underestimation at the highest frequencies, but less than the coarser-resolution models

Bano-Medina et al. (2025)

DIFFUSION MODELS TO GENERATE ENSEMBLES

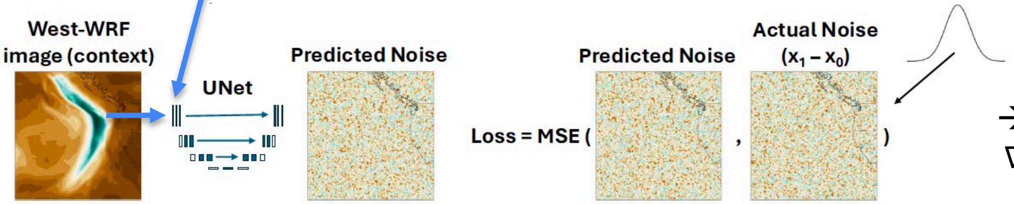
CORE IDEA A diffusion model learns to generate **realistic future weather states** by reversing a process that gradually adds random noise.

TRAINING STEPS



(1) Start from the reference ground-truth, in this case ERA5 (X_0)

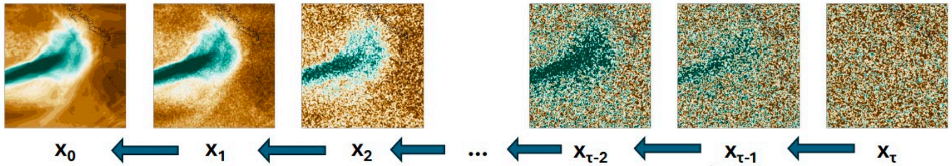
(2) Gradually add noise until the state becomes nearly random (X_0, X_1, $X_{2, \ldots,}$ X_t)



→ This is effectively learning ∇_{ERA5} log p(ERA5 | West-WRF)

(3) Train a neural network to learn how to remove that noise step by step

INFERENCE

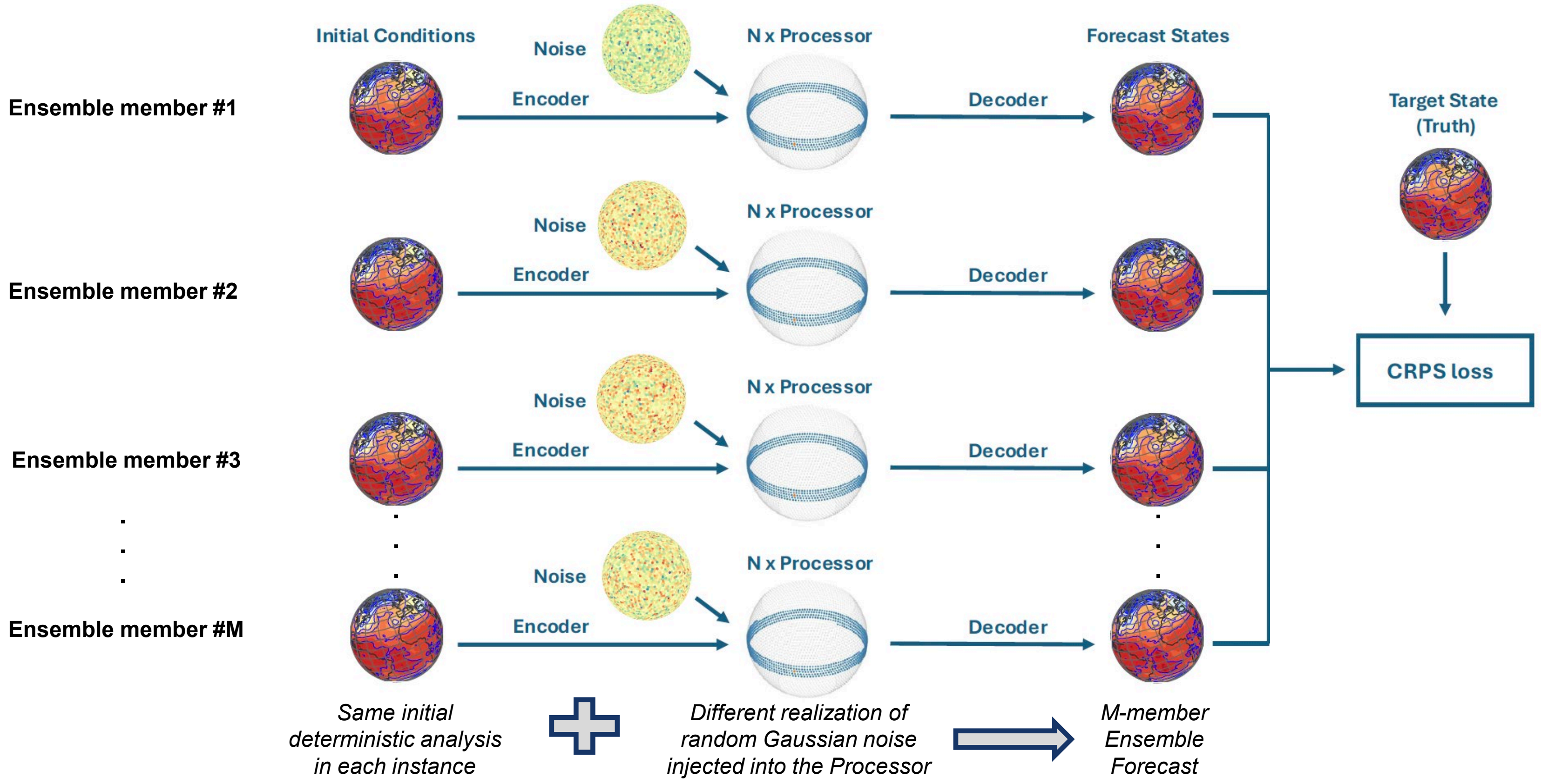


→ The ensemble members will sample the uncertainty of West-WRF prediction, given ERA5 as ground-truth

(4) After training, start from noise and iteratively denoise, **conditioned on the current atmosphere**, to produce a plausible ensemble member

Higgins et al., 2026: Generating a 1000-member ensemble of integrated water vapor transport forecasts with diffusion. *Artificial Intelligence for the Earth Systems, submitted*

ENSEMBLE FORECASTING WITH CRPS-BASED LOSS FUNCTION AT ECMWF



Lang et al., 2026: AIFS-CRPS: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *npj Artificial Intelligence*

EMERGENCE OF AI WEATHER DATA-DRIVEN MODELS

UPPER-AIR VARIABLES

Group	Model	Geopotential (500hPa geopotential RMSE [kg²/m²]) 1	3	5	7	10	Temperature (850hPa temperature RMSE [K]) 1	3	5	7	10	Humidity (700hPa specific humidity RMSE [g/kg]) 1	3	5	7	10	Wind Vector (850hPa wind vector RMSE [m/s]) 1	3	5	7	10
Physical models	IFS HRES	42	135	304	521	801	0.62	1.16	1.82	2.63	3.63	0.55	0.96	1.27	1.53	1.81	1.69	3.30	5.21	7.13	9.16
Physical models	IFS ENS (mean)	42	132	277	439	621	0.65	1.11	1.62	2.17	2.80	0.51	0.84	1.06	1.22	1.38	1.63	2.99	4.45	5.75	6.95
Physical models	ERA5-Forecasts	43	142	316	534	811	0.59	1.19	1.87	2.68	3.66	0.53	1.01	1.33	1.59	1.86	1.63	3.41	5.38	7.27	9.25
ML / hybrid models	Pangu-Weather (oper.)	45	136	300	510	785	0.65	1.09	1.74	2.54	3.55	0.53	0.86	1.17	1.45	1.76	1.71	3.03	4.86	6.77	8.84
ML / hybrid models	GraphCast (oper.)	40	124	276	477	754	0.53	0.93	1.56	2.36	3.40	0.48	0.76	1.03	1.29	1.60	1.48	2.74	4.52	6.42	8.57
ML / hybrid models	GenCast (oper.) (mean)	41	129	274	440	623	0.55	0.96	1.51	2.11	2.76	0.49	0.78	1.00	1.18	1.35	1.54	2.80	4.30	5.67	6.90
ML / hybrid models	Keisler (2022)	66	174	345	544	787	0.81	1.22	1.87	2.63	3.55	0.65	0.94	1.19	1.41	1.65	2.27	3.51	5.18	6.87	8.64
ML / hybrid models	Pangu-Weather	44	133	294	501	778	0.62	1.05	1.71	2.51	3.54	0.53	0.88	1.19	1.47	1.79	1.66	3.01	4.83	6.73	8.81
ML / hybrid models	GraphCast	39	124	274	467	731	0.51	0.94	1.55	2.33	3.36	0.47	0.79	1.06	1.30	1.59	1.42	2.76	4.45	6.23	8.20
ML / hybrid models	FuXi	40	125	277	433	631	0.54	0.97	1.59	2.14	2.91						1.47	2.80	4.51	5.66	7.04
ML / hybrid models	NeuralGCM 0.7	37	115	267	469	751	0.54	0.97	1.58	2.38	3.42	0.48	0.83	1.12	1.40	1.71	1.49	2.81	4.59	6.51	8.66
ML / hybrid models	NeuralGCM ENS (mean)	43	126	266	424	606	0.65	1.02	1.53	2.10	2.75	0.54	0.81	1.02	1.19	1.37	1.76	2.89	4.29	5.60	6.84
ML / hybrid models	GenCast (mean)	39	123	262	420	606	0.54	0.95	1.48	2.07	2.73	0.49	0.80	1.02	1.19	1.37	1.49	2.78	4.23	5.55	6.83
ML / hybrid models	Stormer ENS (mean)	45	134	285	460	665	0.60	1.01	1.59	2.24	3.00	0.52	0.82	1.05	1.25	1.44	1.65	2.91	4.50	5.95	7.30
ML / hybrid models	Excarta (HEAL-ViT)	38	117	271	476	763	0.53	0.95	1.60	2.41	3.49										
ML / hybrid models	ArchesWeather-Mx4	41	132	288	473	693	0.59	1.00	1.62	2.33	3.11	0.51	0.82	1.09	1.32	1.54	1.68	2.94	4.62	6.21	7.66
ML / hybrid models	ArchesWeatherGen (mean)	41	130	272	428	610	0.59	0.98	1.51	2.09	2.75	0.51	0.80	1.01	1.19	1.37	1.66	2.87	4.30	5.59	6.84
	Climatology	820	820	820	820	820	3.44	3.44	3.44	3.44	3.44	1.59	1.59	1.59	1.59	1.59	7.90	7.90	7.90	7.90	7.90

Lead time [days]   Lead time [days]   Lead time [days]   Lead time [days]

−50 −20 −10 −5 −2 −1 1 2 5 10 20 50

Better ⟵ % difference in RMSE vs IFS HRES ⟶ Worse

Scores for physics-based vs AI models

- Evaluated against ECMWF's IFS HRES for lead times of 1-10 days
- Testing period: the year of 2020
- Blues: performance better than IFS
- Reds: performance worse than IFS

Source: Google's WeatherBench project

EMERGENCE OF AI WEATHER DATA-DRIVEN MODELS

SURFACE VARIABLES

Group	Model	2m Temperature RMSE [K] 1	3	5	7	10	Surface Pressure RMSE [Pa] 1	3	5	7	10	10m Wind Speed RMSE [m/s] 1	3	5	7	10	Precipitation 24h precipitation SEEPS 1	3	5	7	10
Physical models	IFS HRES	0.51	0.87	1.33	1.88	2.56	60	149	309	506	748	0.70	1.27	1.92	2.49	3.00	0.20	0.33	0.46	0.60	0.76
Physical models	IFS ENS (mean)	0.56	0.86	1.20	1.58	2.00	62	146	280	423	575	0.67	1.14	1.61	1.96	2.25	0.21	0.35	0.49	0.63	0.77
Physical models	ERA5-Forecasts	0.71	1.06	1.51	2.06	2.70	57	155	321	519	755	0.65	1.30	1.96	2.52	3.01					
ML / hybrid models	Pangu-Weather (oper.)	0.58	0.95	1.40	1.95	2.64	59	147	302	493	732	0.72	1.22	1.84	2.41	2.95					
ML / hybrid models	GraphCast (oper.)	0.44	0.75	1.18	1.73	2.43	52	133	279	465	707	0.65	1.13	1.75	2.34	2.90	0.21	0.32	0.43	0.56	0.72
ML / hybrid models	GenCast (oper.) (mean)	0.44	0.75	1.13	1.54	1.98	53	137	275	424	578	0.60	1.06	1.55	1.93	2.22	0.20	0.31	0.44	0.58	0.74
ML / hybrid models	Pangu-Weather	0.56	0.91	1.39	1.97	2.68	55	143	297	486	725	0.65	1.17	1.80	2.37	2.92					
ML / hybrid models	GraphCast	0.51	0.82	1.25	1.81	2.56	48	131	276	453	682	0.60	1.13	1.74	2.30	2.84	0.21	0.50	0.60	0.69	0.81
ML / hybrid models	FuXi	0.53	0.84	1.27	1.65	2.17	49	132	277	416	584	0.60	1.11	1.72	2.47	3.02	0.14	0.31	0.44	0.66	0.79
ML / hybrid models	GenCast (mean)	0.49	0.80	1.16	1.57	2.02	48	131	264	406	563	0.56	1.03	1.51	1.88	2.20	0.14	0.29	0.43	0.57	0.73
ML / hybrid models	Stormer ENS (mean)	0.66	0.93	1.30	1.74	2.23	55	142	286	444	619	0.72	1.23	1.86	2.43	3.07					
ML / hybrid models	Excarta (HEAL-ViT)	0.55	0.84	1.27	1.83	2.56						0.56	1.06	1.70	2.30	2.90					
ML / hybrid models	ArchesWeather-Mx4	0.51	0.83	1.27	1.77	2.32	52	140	289	457	641	0.77	1.24	1.84	2.38	2.95					
ML / hybrid models	ArchesWeatherGen (mean)	0.54	0.83	1.20	1.60	2.04	51	137	272	413	567	0.69	1.10	1.55	1.91	2.21					
	Climatology	2.61	2.61	2.61	2.61	2.61	725	725	725	725	725	2.43	2.43	2.43	2.43	2.43	0.95	0.95	0.95	0.95	0.95

Lead time [days]

−50 −20 −10 −5 −2 −1 1 2 5 10 20 50

Better ⟵ % difference in RMSE/SEEPS vs IFS HRES ⟶ Worse

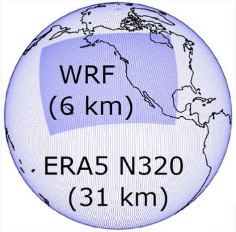
Scores for physics-based vs AI models

- Evaluated against ECMWF’s IFS HRES for lead times of 1-10 days
- Testing period: the year of 2020
- Blues: performance better than IFS
- Reds: performance worse than IFS

Source: Google’s WeatherBench project

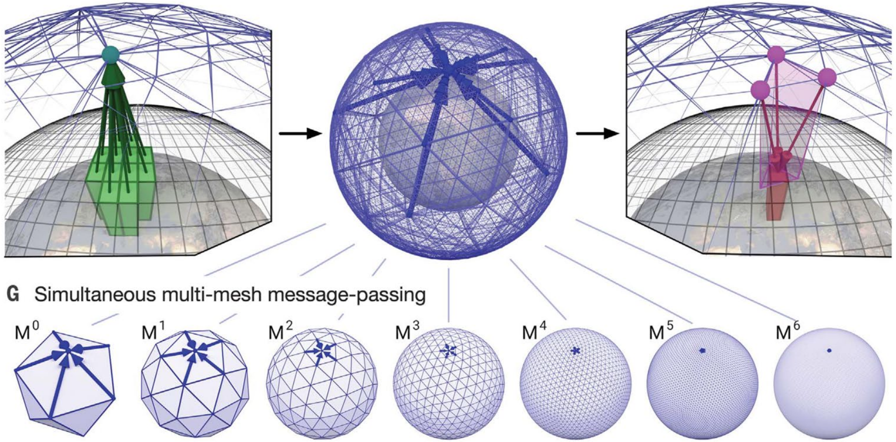
TASK 1 – ARCHITECTURE OF THE GLOBAL AND REGIONAL AI MODELS

INPUT FEATURS
several meteorological fields on the original grid

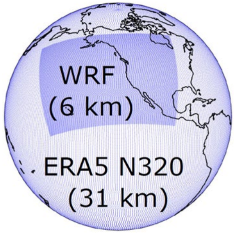


@ t

DECODER **PROCESSOR** **DECODER**



AI MODEL FORECAST

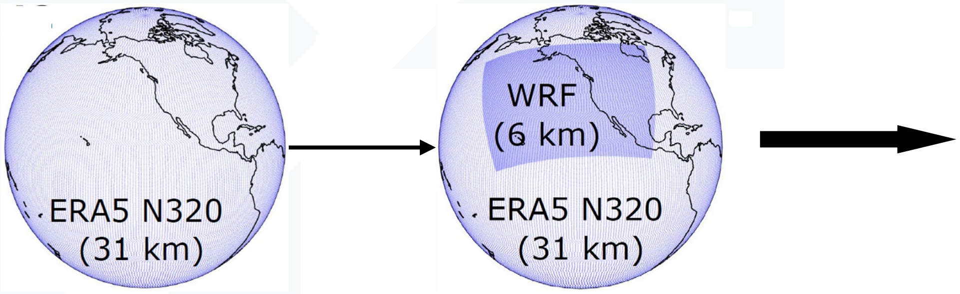


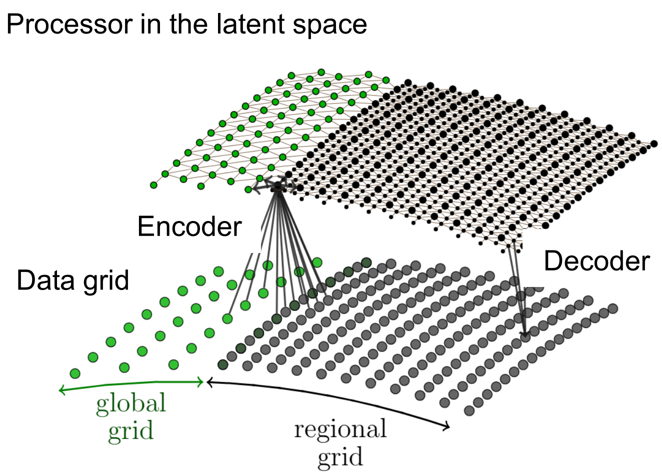
@ t+1

Graph neural network (GNN)-based global weather prediction model (Lam et al. 2023)

TASK 1 – GRID REFINEMENT

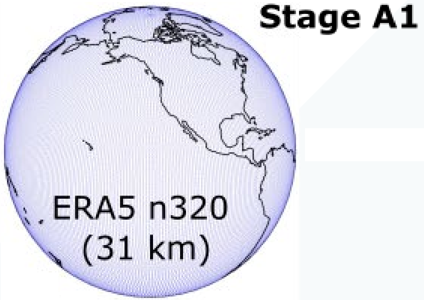
REGIONAL REFINEMENT





Nipen et al. (2025)

TASK 1 – TRAINING FRAMEWORK OF THE GLOBAL AND REGIONAL MODELS

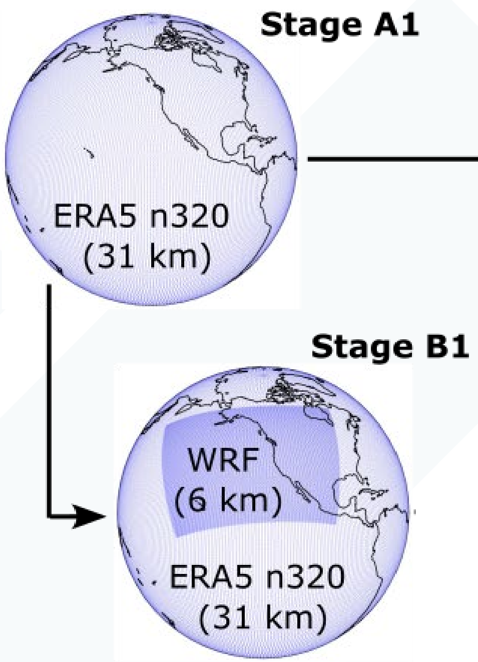


- STAGE A1: Global model. Optimized to forecast the next 6-hr of weather. ✅

TASK 1 – TRAINING FRAMEWORK OF THE GLOBAL AND REGIONAL MODELS

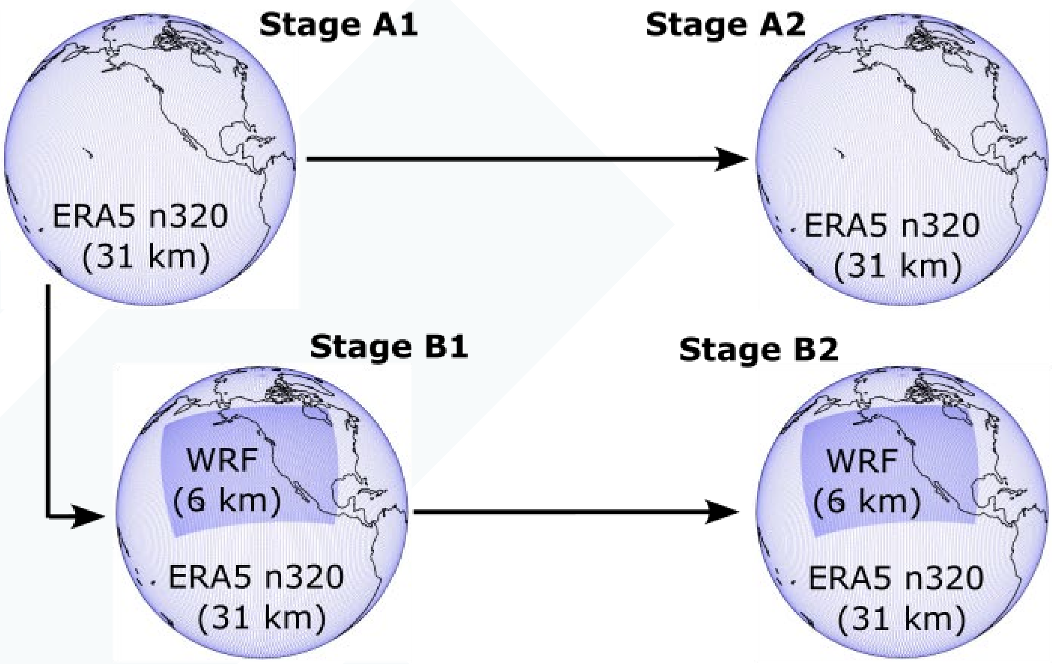
- STAGE A1: Global model. Optimized to forecast the next 6-hr of weather. ✅
- STAGE A2: Global model. Optimized to forecast the next 72-hr of weather → for stability in long runs. ✅

TASK 1 – TRAINING FRAMEWORK OF THE GLOBAL AND REGIONAL MODELS



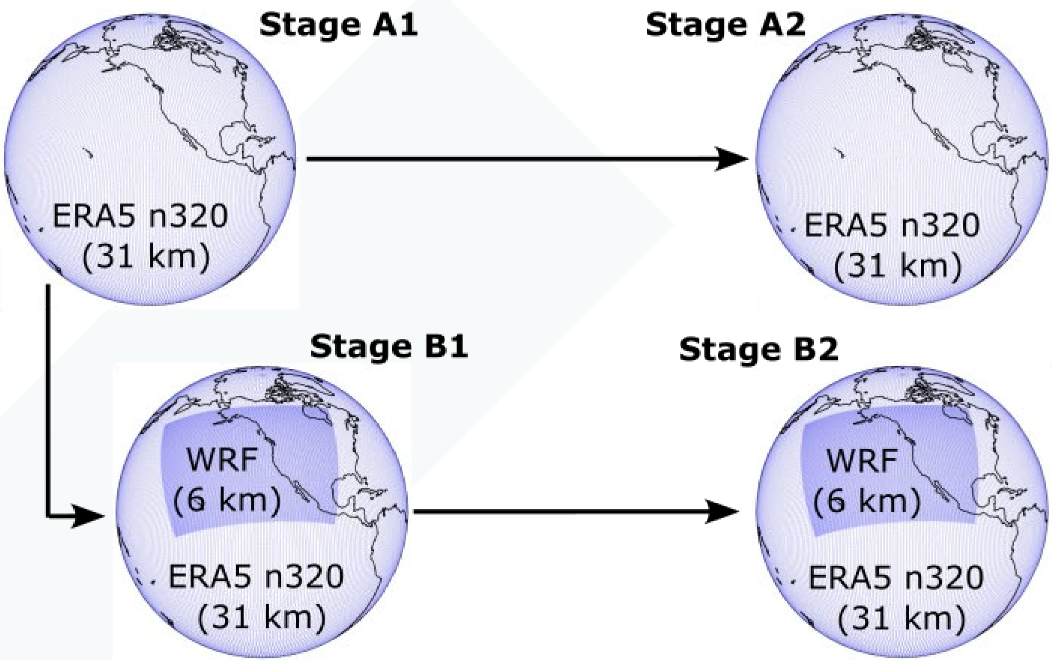
- STAGE A1: Global model. Optimized to forecast the next 6-hr of weather. ✅
- STAGE A2: Global model. Optimized to forecast the next 72-hr of weather → for stability in long runs. ✅
- STAGE B1: Stretched-grid model. Optimized to forecast the next 6-hr of weather.

TASK 1 – TRAINING FRAMEWORK OF THE GLOBAL AND REGIONAL MODELS



- STAGE A1: Global model. Optimized to forecast the next 6-hr of weather. ✅
- STAGE A2: Global model. Optimized to forecast the next 72-hr of weather → for stability in long runs. ✅
- STAGE B1: Stretched-grid model. Optimized to forecast the next 6-hr of weather.
- STAGE B2: Stretched-grid model. Optimized to forecast the next 72-hr of weather → for stability in long runs.

TASK 1 – TRAINING FRAMEWORK OF THE GLOBAL AND REGIONAL MODELS



Notes

87 variables

13 vertical levels

6-hr temporal resolution

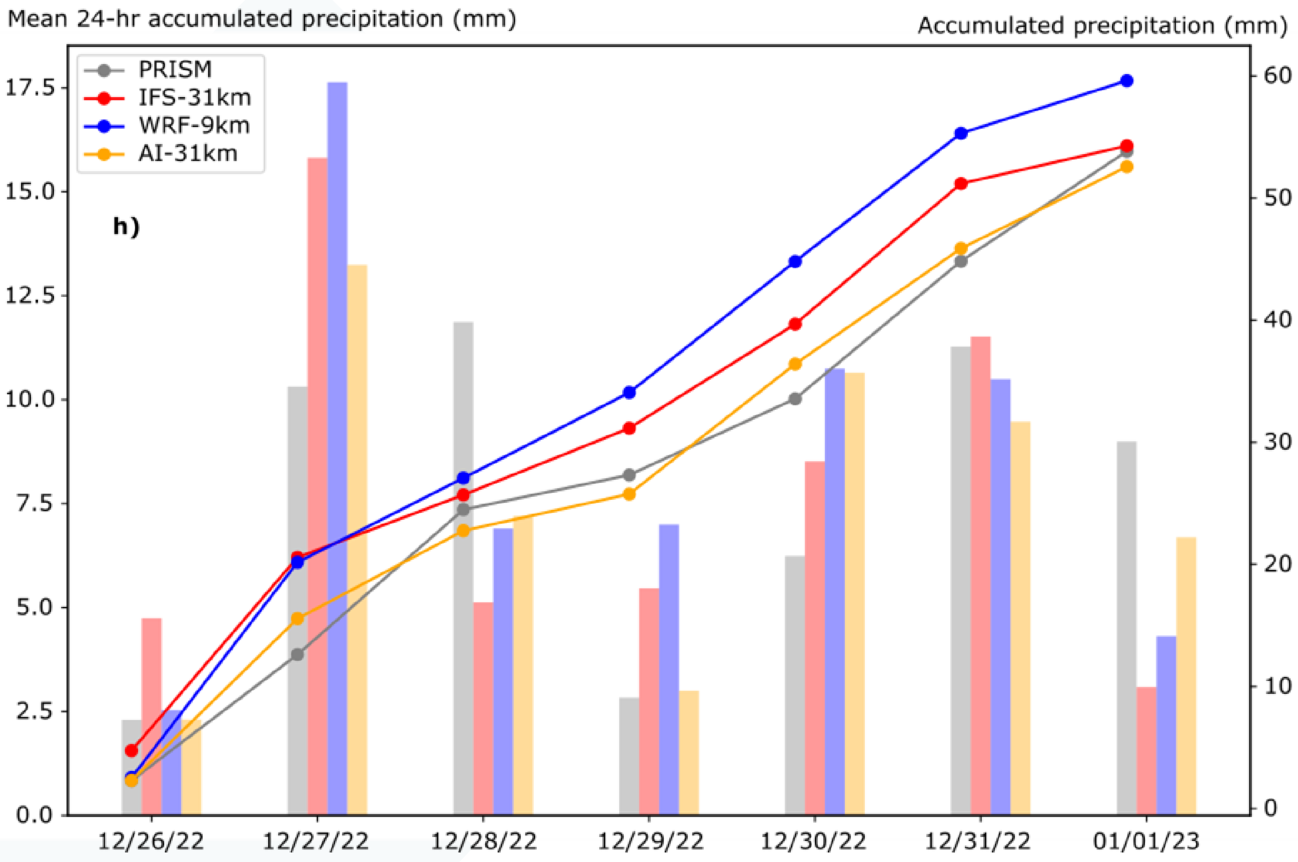
Weeks to train…

Minutes to predict!

- STAGE A1: Global model. Optimized to forecast the next 6-hr of weather. ✅
- STAGE A2: Global model. Optimized to forecast the next 72-hr of weather → for stability in long runs. ✅
- STAGE B1: Stretched-grid model. Optimized to forecast the next 6-hr of weather.
- STAGE B2: Stretched-grid model. Optimized to forecast the next 72-hr of weather → for stability in long runs.

TASK 1 – CASE STUDY: NEW YEAR'S EVE ATMOSPHERIC RIVER OF 2022

__Fig.__ Temporal series of 24-hr accumulated precipitation for the New Year's Eve Atmospheric River of 2022. Solid lines represent the accumulated precipitation throughout the series.



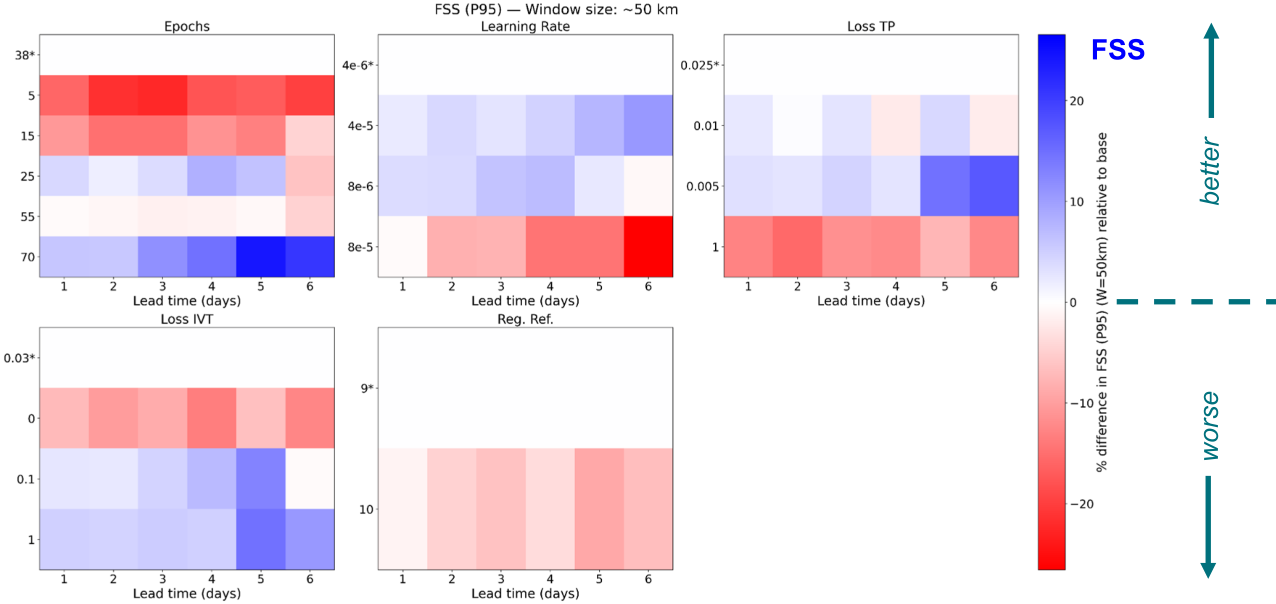
TASK 1 – 2 KM MODEL, SYSTEMATIC HYPERPARAMETER EXPLORATION

Hyperparameter Sensitivity Analysis

- Winter 2020–2023
- Spatial Resolution: 4 km

Experimenting with:

- Climatological Precipitation percentile **p95**
- Evaluation metrics: **FSS**, **RMSE, CC**
- Neighborhood window size: **50 km**



* (y-axis): base model hyperparameter

Rouzegari et al. (2026, in preparation)

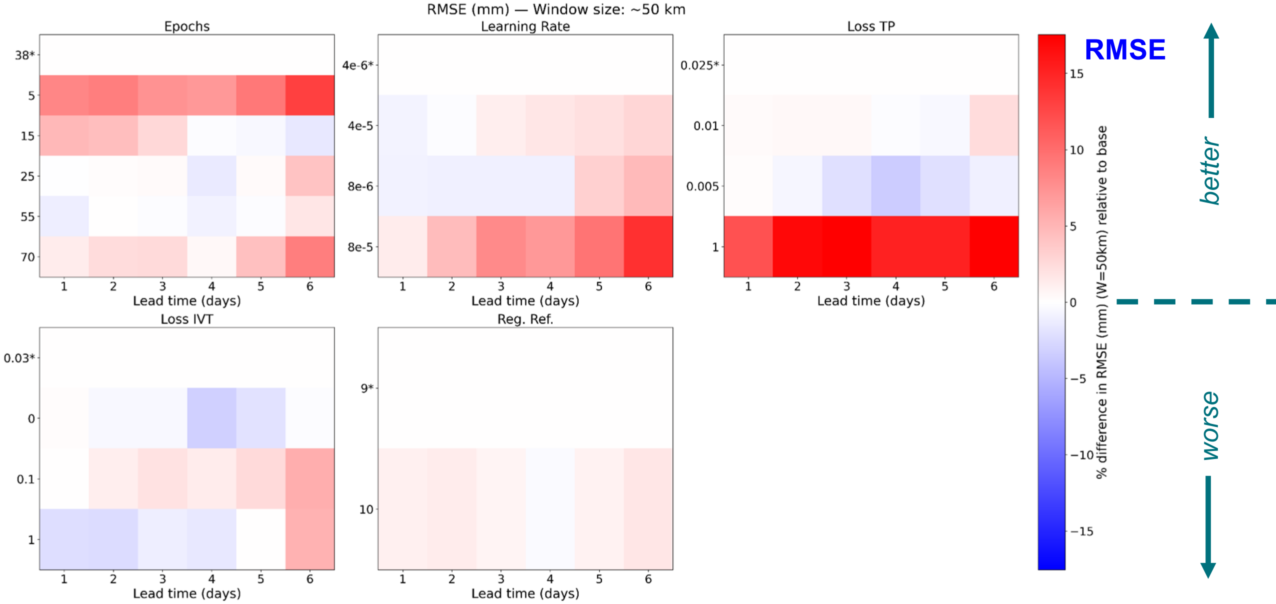
TASK 1 – 2 KM MODEL, SYSTEMATIC HYPERPARAMETER EXPLORATION

Hyperparameter Sensitivity Analysis

- Winter 2020–2023
- Spatial Resolution: 4 km

Experimenting with:

- Climatological Precipitation percentile **p95**
- Evaluation metrics: **FSS, RMSE, CC**
- Neighborhood window size: **50 km**
- Max value within the sliding window at every grid point



* (y-axis): base model hyperparameter

Rouzegari et al. (2026, in preparation)

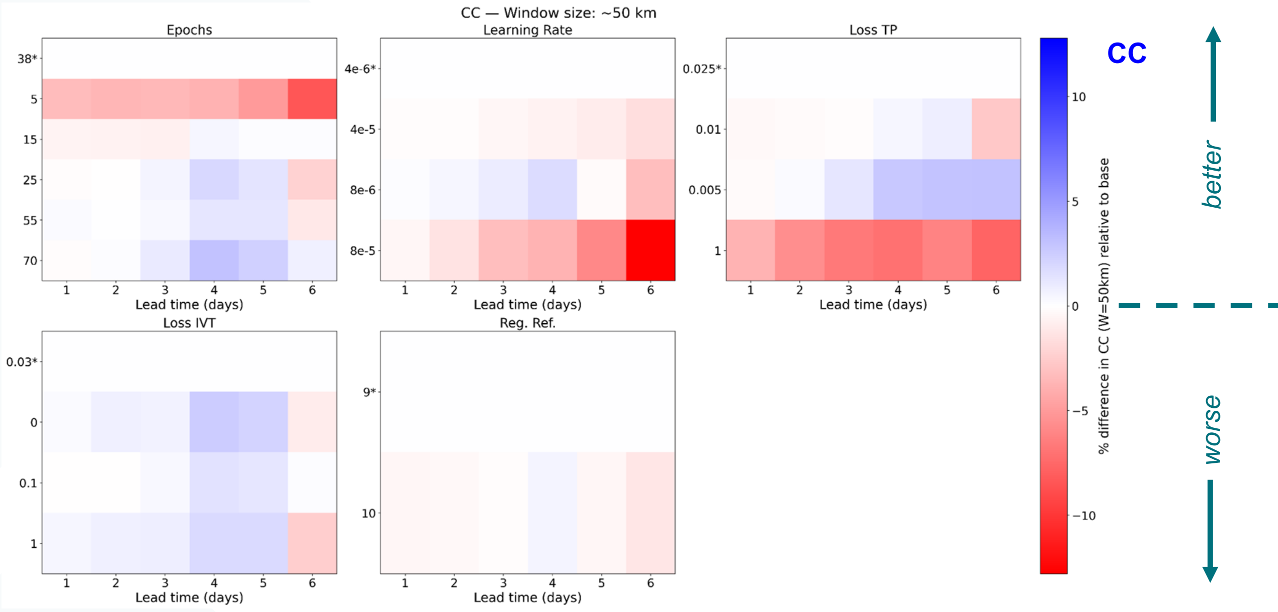
TASK 1 – 2 KM MODEL, SYSTEMATIC HYPERPARAMETER EXPLORATION

Hyperparameter Sensitivity Analysis

- Winter 2020–2023
- Spatial Resolution: 4 km

Experimenting with:

- Climatological Precipitation percentile **p95**
- Evaluation metrics: **FSS, RMSE, CC**
- Neighborhood window size: **50 km**
- Max value within the sliding window at every grid point



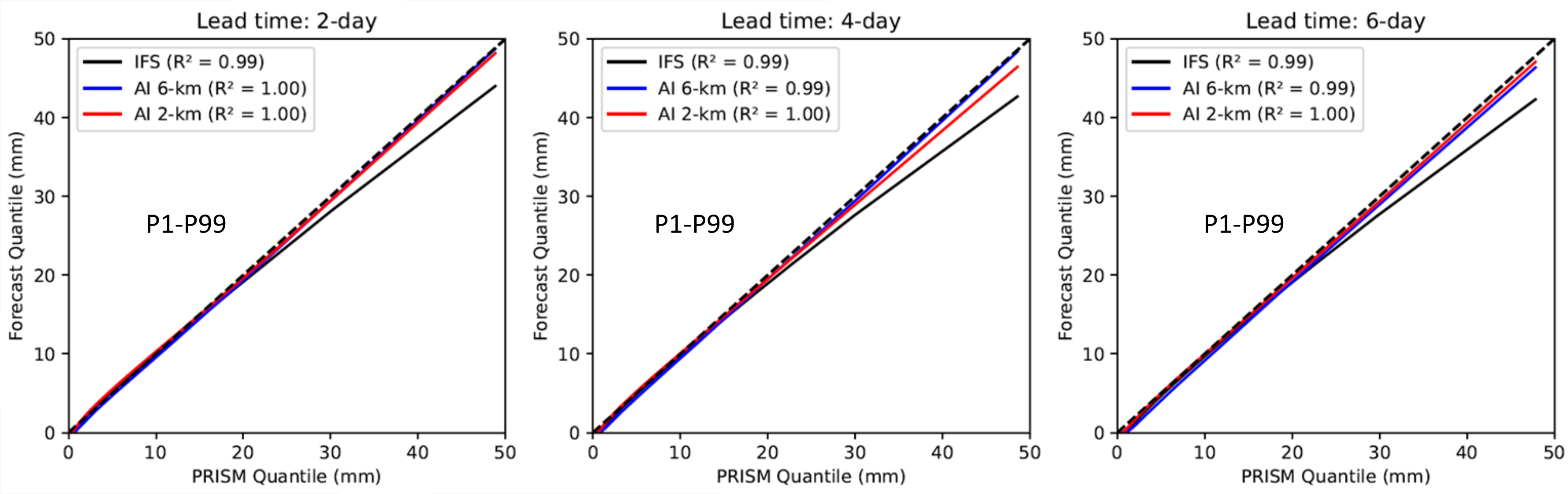
* (y-axis): base model hyperparameter

Rouzegari et al. (2026, in preparation)

TASK 1 – 2 KM MODEL PRELIMINARY EVALUATION

Model Verification

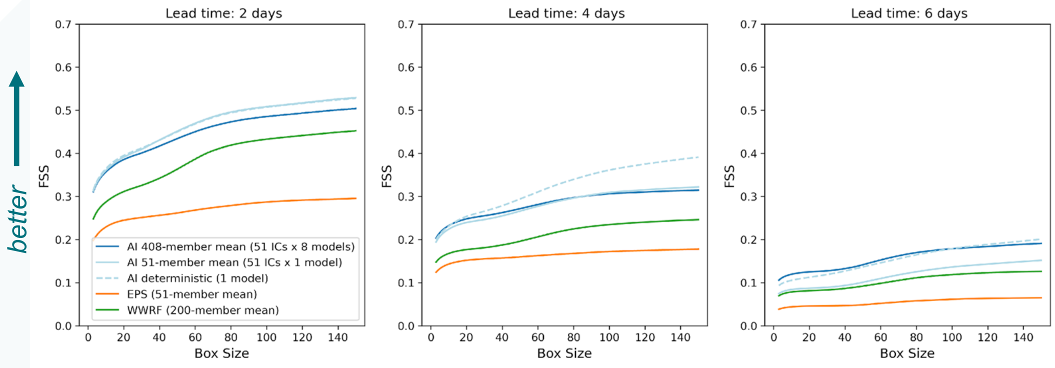
- Winter 2020–2023
- Spatial Resolution: 4 km



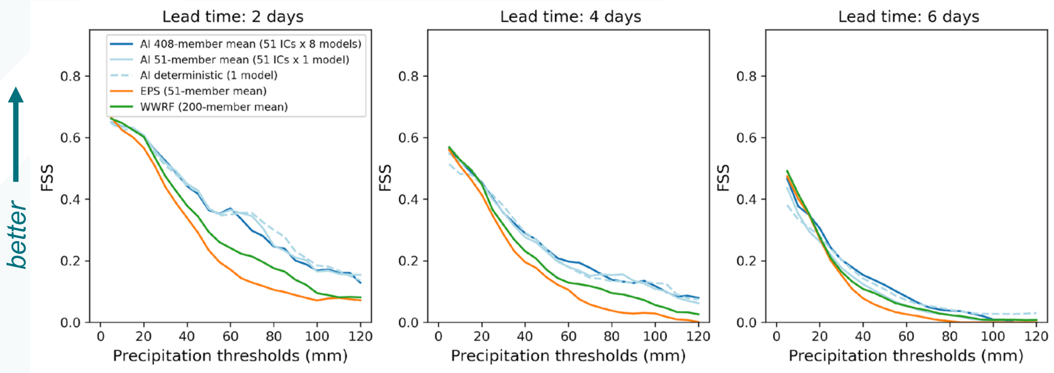
Rouzegari et al. (2026, in preparation)

TASK 2 – PRELIMINARY ASSESSEMENT OF AI ENSEMBLE MEAN

Threshold = 50 mm



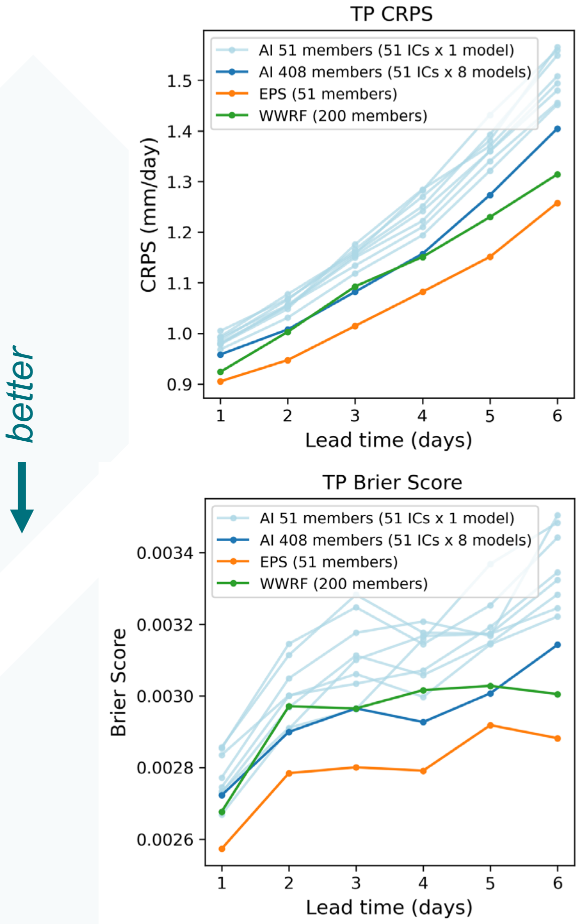
Box size = 48 km (12 grid points)



- Metric: Fractional Skill Score (FSS) for 24-h accumulated precipitation over the Western US
- Testing period: DJF 2023–2024 and 2024–2025
- The AI models outperform dynamical models for higher precipitation thresholds (>20 mm)

Nie et al. (2026, in preparation)

TASK 2 – PRELIMINARY ASSESSEMENT OF AI ENSEMBLE

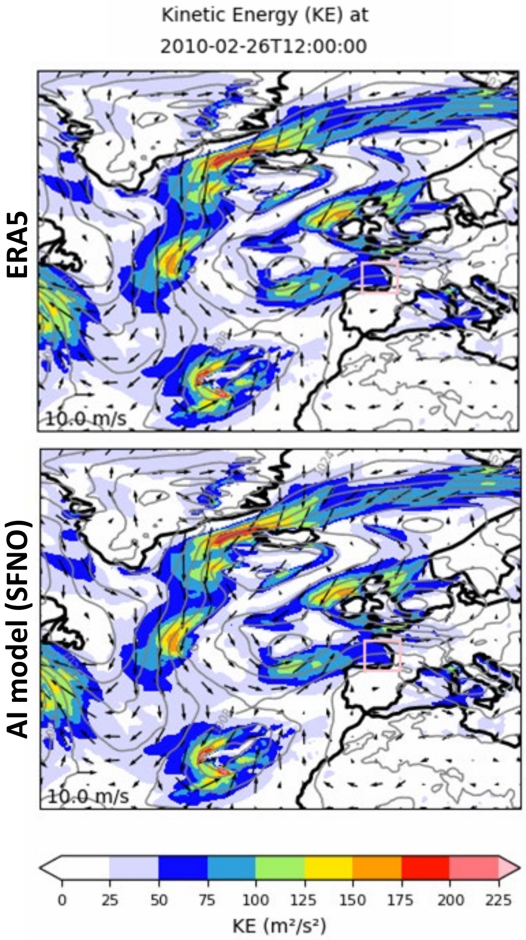


- Metric: Brier score (threshold = 50 mm) and CRPS for 24-h accumulated precipitation over the Western US
- Testing period: DJF 2023–2024 and 2024–2025
- The 408-member AI ensemble is comparable to West-WRF
- The 408-member AI ensemble generated by eight models outperforms the 51-member AI ensembles generated by one model

Nie et al. (2026, in preparation)

TASK 3 – ARE AI DATA-DRIVEN MODELS LEARNING PHYSICS?

- Are AI models learning physically-meaningful spatio-temporal links between atmospheric processes informing forecast evolution?
  - Evidence of the physical realism of AI models and/or diagnose model flaws and potential improvements.
  - Open new pathways of exploration and application:
    - Initial condition sensitivity and predictability studies (e.g., Vonich and Hakim, 2024)
    - Observation-network design (AR Recon, Ralph et al., 2020)
    - Process-based studies



TASK 3 – ARE AI DATA-DRIVEN MODELS LEARNING PHYSICS?

To answer the question: compare physics-based and AI-based sensitivity analysis

Sensitivity analysis allow one to examine how a target forecast metric/variable of interest (K), e.g., Kinetic Energy (KE) over Western Europe, responds to (infinitesimal) perturbations to the atmospheric patterns at the initial time (Xi) at an upstream location, i.e., the gradient **$\partial K/\partial Xi$**

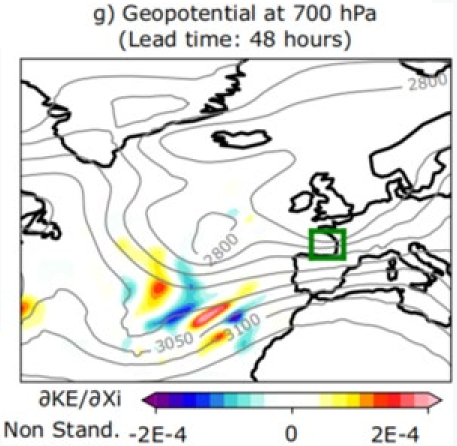
- Physics-based models → Adjoint model
  - Assumes linear perturbation growth
  - Computationally expensive

- AI-based models → Backpropagation
  - Does not assumes linear perturbation growth → Might be useful for long lead times (> 5 days)
  - Computationally inexpensive

Baño-Medina, et al. (2025b)

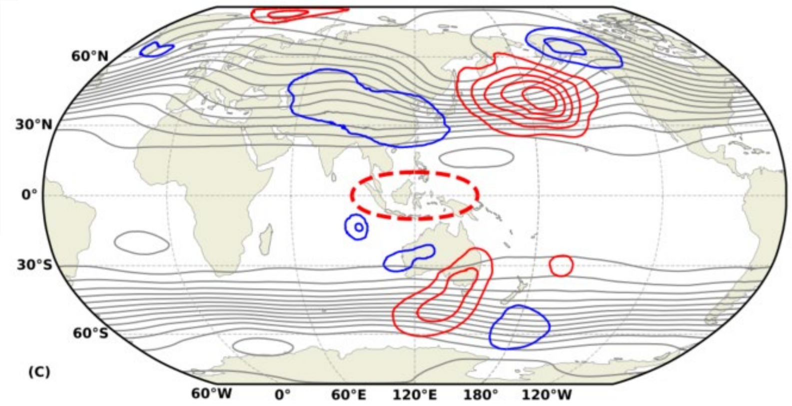
TASK 3 – ARE AI DATA-DRIVEN MODELS LEARNING PHYSICS?

Initial condition sensitivity for cyclone Xynthia based on SFNO



Baño-Medina et al. (2025a)

Response in 500-hPa geopotential height to tropical heating based on Pangu-Weather



Hakim and Masanam (2024)

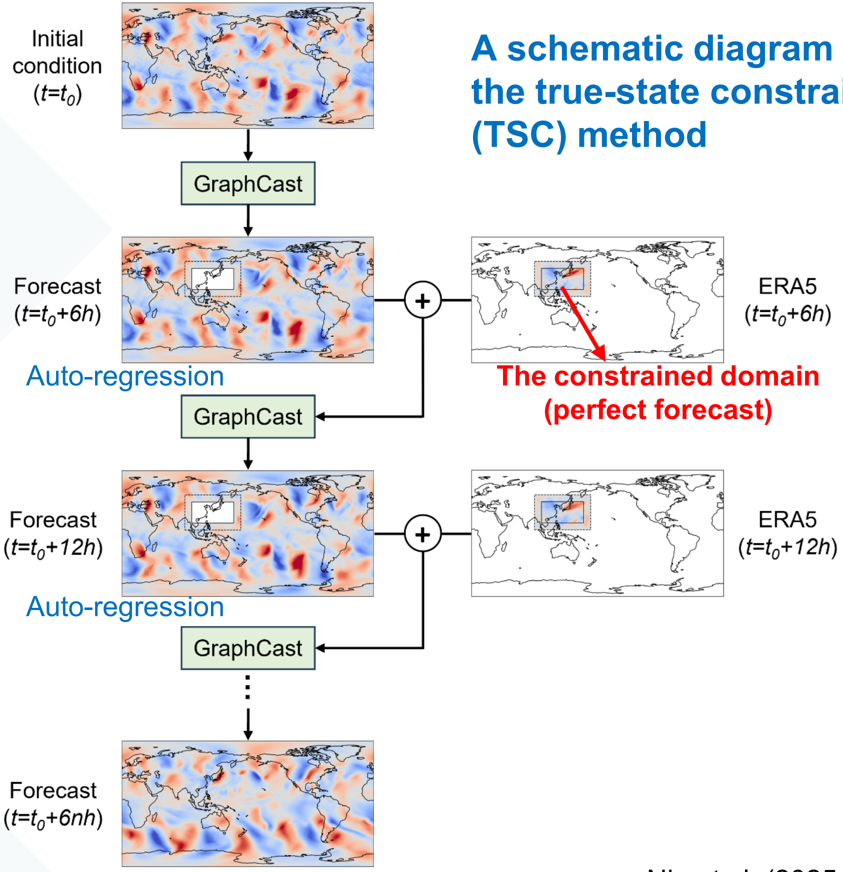
AI weather models have been applied to

- Initial condition sensitivity analysis (Baño-Medina et al., 2025)
- Climate attribution (Jiménez-Esteve et al., 2025)
- Idealized dynamical experiments (Hakim and Masanam, 2024)

TASK 3 – AI-BASED NUDGING FOR PROCESS-BASED STUDIES

After each auto-regression step (i.e., after the generation of a forecast lead time), the forecasted atmospheric states are replaced by concurrent ERA5 data within the constrained region before being sent to the model for the following step. The model runs freely elsewhere.

The solid box denotes the constrained region. The area between the solid and dashed boxes is the buffer zone.

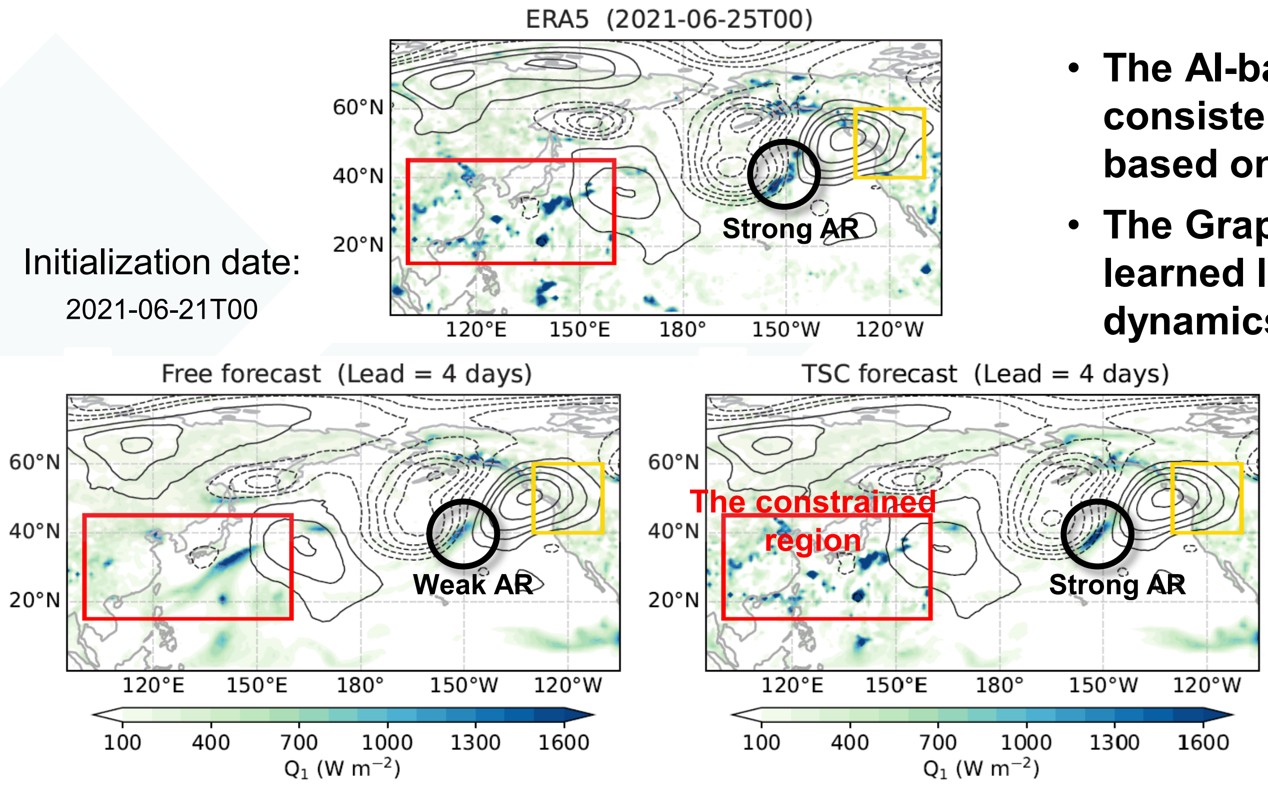


A schematic diagram of the true-state constraint (TSC) method

Nie et al. (2025, submitted)

TASK 3 – AI-BASED NUDGING FOR PROCESS-BASED STUDIES

Initialization date:
2021-06-21T00



- **The AI-based results are consistent with previous studies based on physics-based models.**
- **The GraphCast model has learned large-scale atmospheric dynamics.**

- Contours: Z500 anomalies
- Shading: apparent heat source (Q_1)

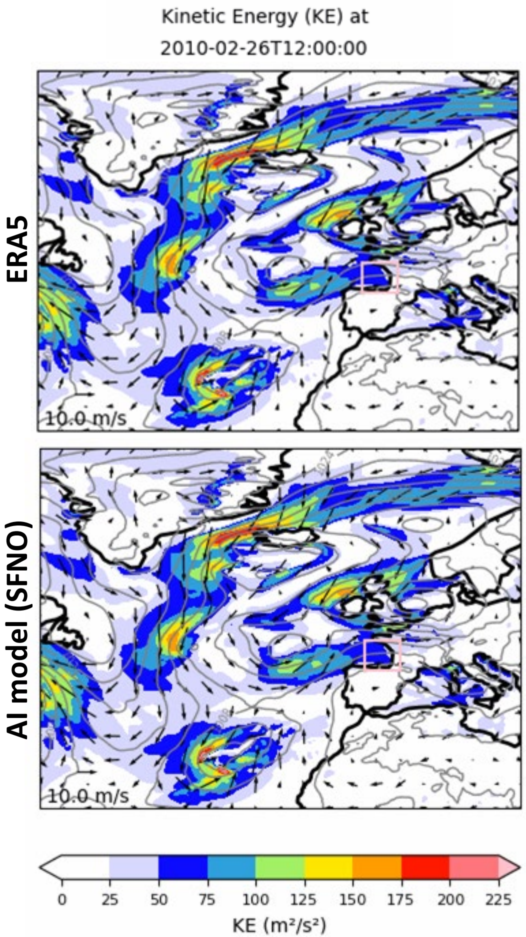
Diabatic heating: $Q_1 = c_p\left(\frac{\partial T}{\partial t} + \mathbf{V}\cdot\nabla T - \omega\sigma\right)$ (Yanai et al. 1973)

Nie et al. (2025, submitted)

TASK 3 – ARE AI DATA-DRIVEN MODELS LEARNING PHYSICS?

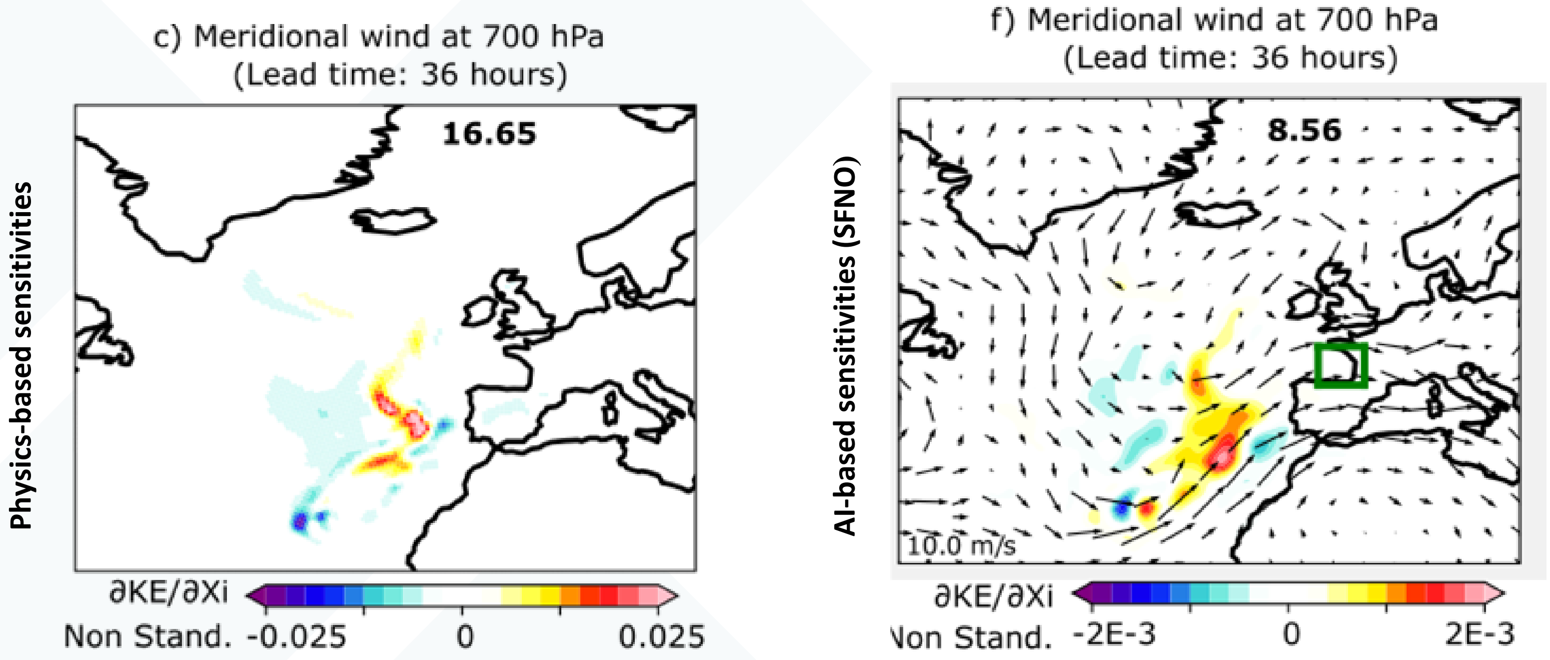
Cyclone Xynthia:

- Developed South at the Azores Islands on February 2010.
- It intensified reaching the coast of Portugal, then the North of Spain and Western France in 36 hours.
- Physics-based sensitivity studies indicate that Xynthia developed due to a filament of moisture within an atmospheric river (Doyle et al. 2014).



TASK 3 – RESULTS

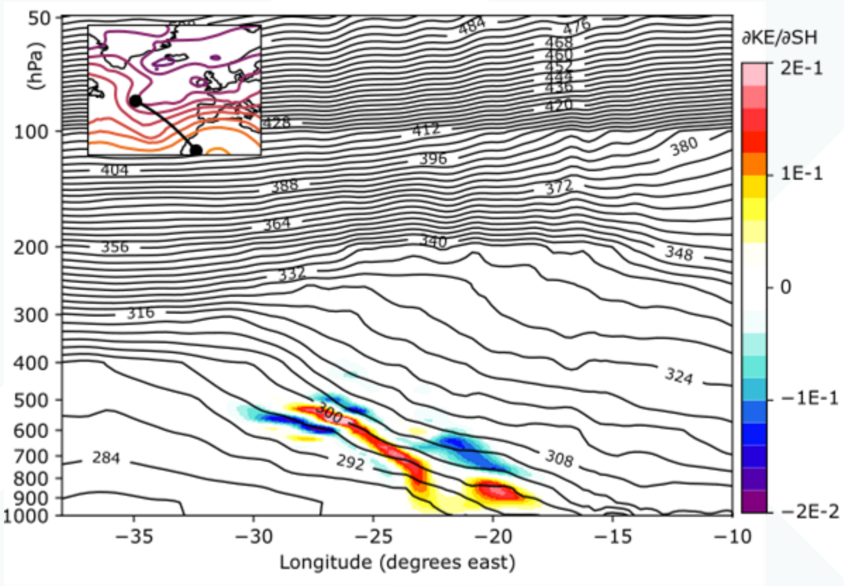
Hypothesis If physics-based and AI-based sensitivities are comparable then AI models are showing evidence of physical realism.



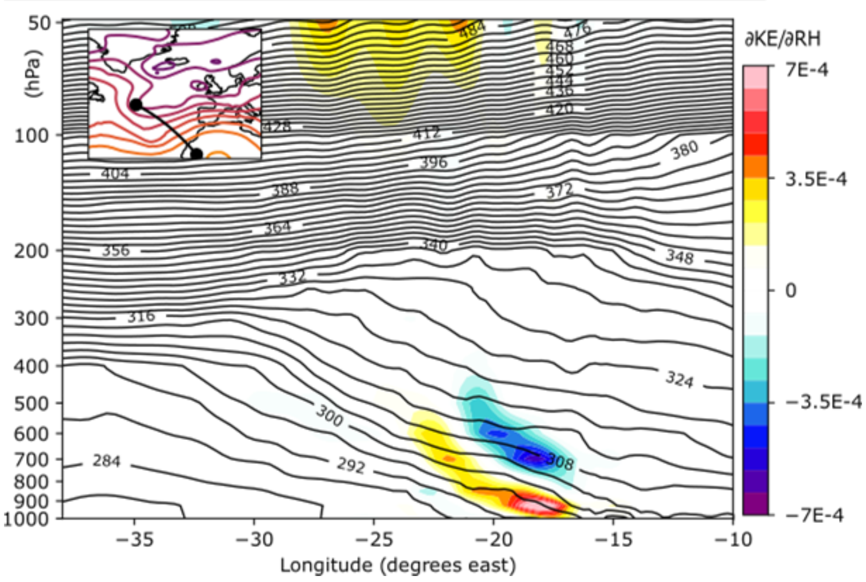
TASK 3 – RESULTS

Hypothesis If physics-based and AI-based sensitivities are comparable then AI models are showing evidence of physical realism.

Physics-based sensitivities

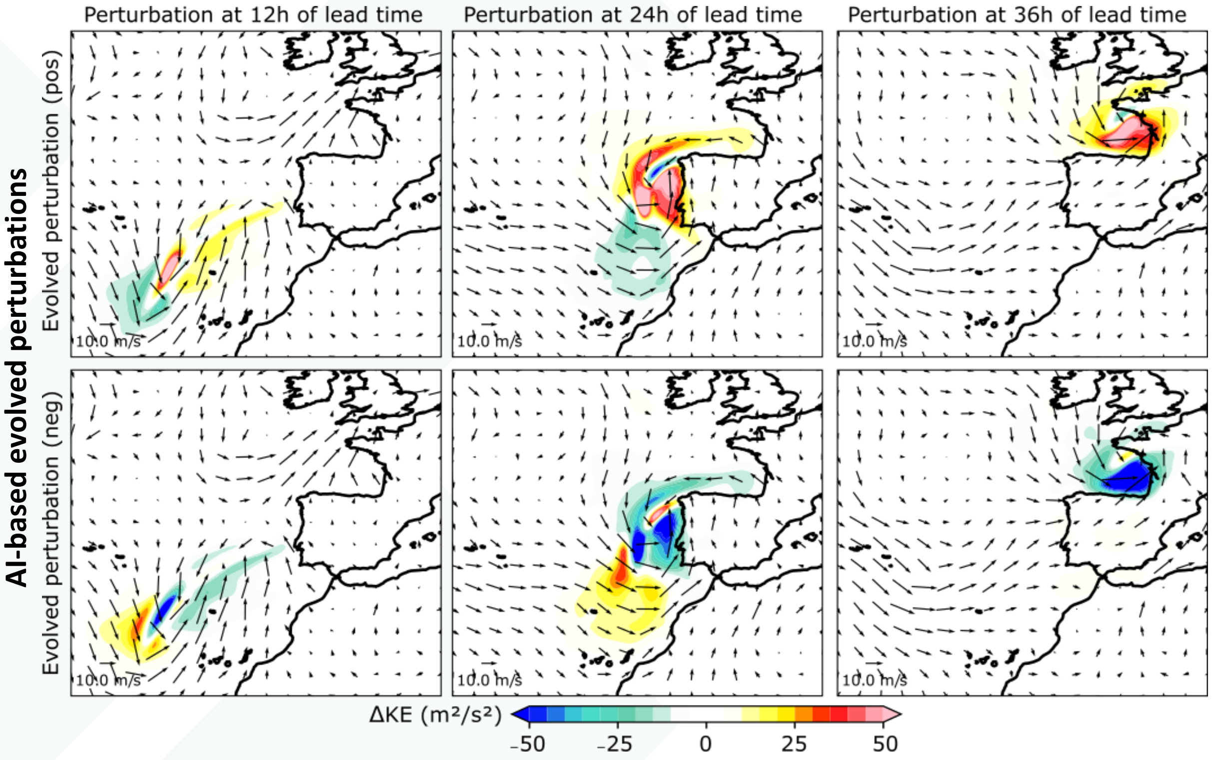


AI-based sensitivities (SFNO)



TASK 3 – EVOLVED PERTURBATIONS

Sensitivity fields are scaled to magnitudes consistent with estimates of initial condition uncertainty, and these perturbation fields are added to the variables at initial time and evolved with the AI model



TASK 3 – RESULTS

- **<u>HYPOTHESIS:</u> If physics-based and AI-based sensitivities are comparable then AI models are showing evidence of physical realism.**