Instituto de Física de Cantabria

CSIC Consejo Superior de Investigaciones Científicas

UC Universidad de Cantabria

WCRP CORDEX Coordinated Regional Climate Downscaling Experiment

Introducing eXplainable Artificial Intelligence to assess Deep Learning models for Statistical Downscaling

Jose González-Abad
gonzabad@ifca.unican.es

Jorge Baño-Medina
bmedina@ifca.unican.es

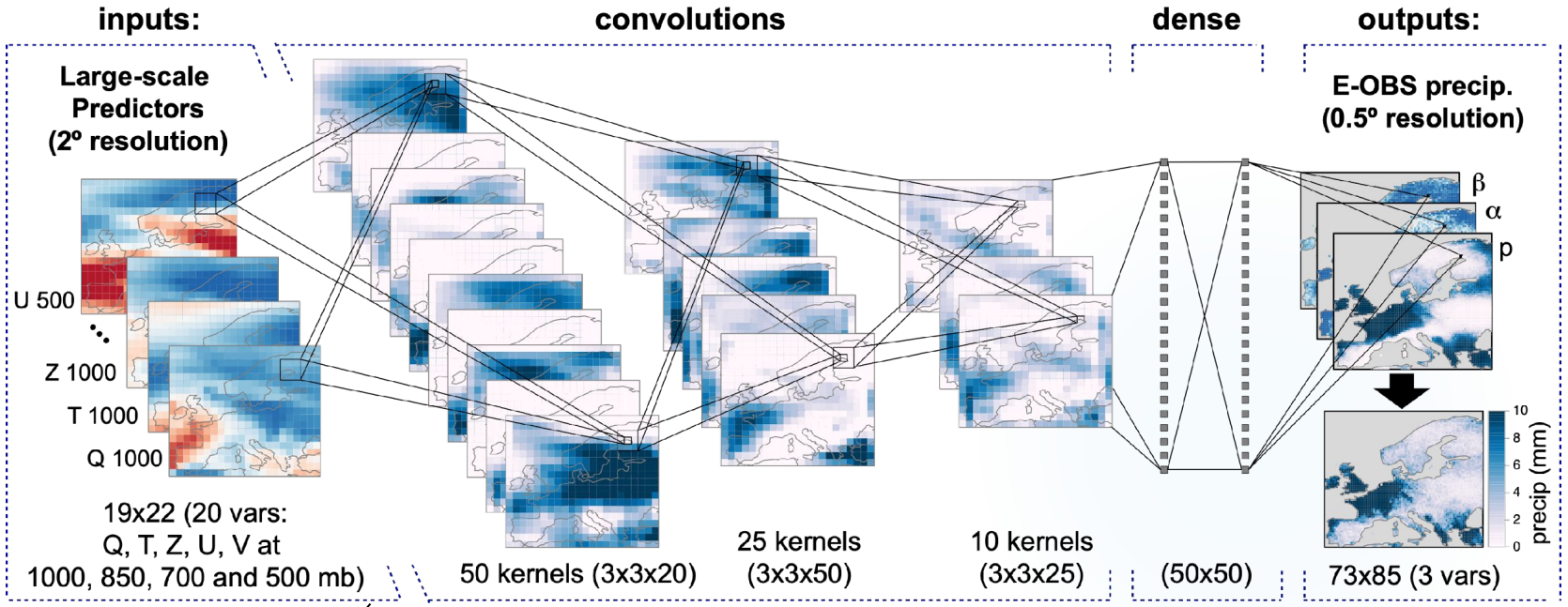
José Manuel Gutiérrez
gutierjm@ifca.unican.es

ICRC-CORDEX 2023:
Statistical Methods/Machine Learning techniques for RCM

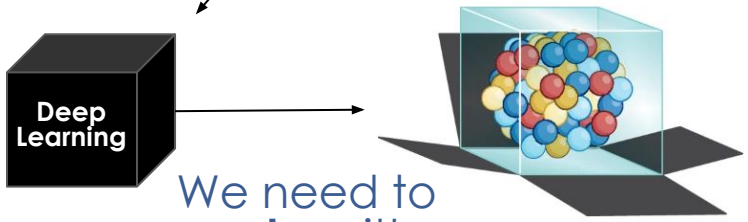
Perfect Prognosis downscaling

Source: Esterhuizen, J. A., Goldsmith, B. R., & Linic, S. (2022). Interpretable machine learning for knowledge generation in heterogeneous catalysis. Nature Catalysis, 1-10.

Deep Learning (DL) has recently emerged as a promising Perfect Prognosis (PP) technique



However, its **black box** nature makes it **difficult** to gain a **comprehensive** understanding of their inner functioning, **particularly for downscaling climate change projections.**

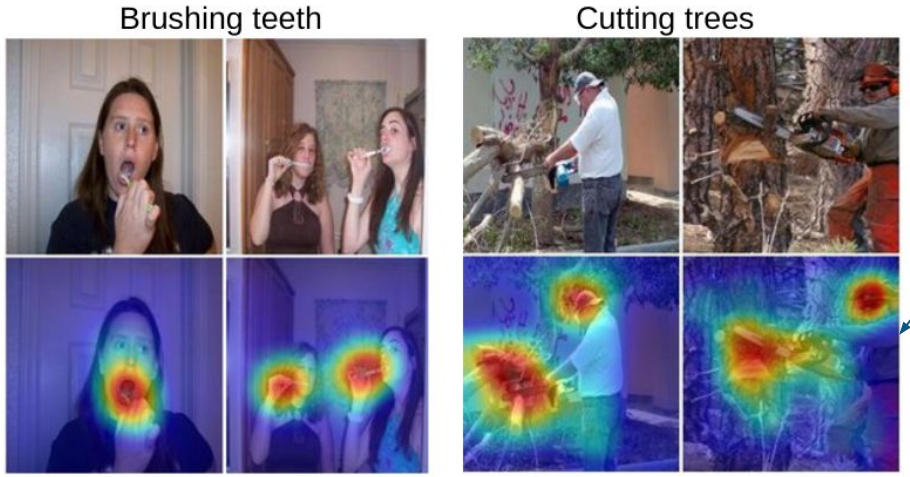


We need to **unbox** it!

Can we **trust** these models?

Interpretability techniques

Interpretability techniques emerged in the computer vision field to **explain the functioning and results of deep learning** models



Source: Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016

Saliency maps assign to each feature a **relevance score** representing its influence on the computed prediction

For downscaling: features ⇒ gridpoints

Relevance *R* of feature *x_i* can be computed as follows:

$$R(x_i) = \frac{\partial f(\mathbf{x})}{\partial x_i}$$

Sensitivity of the prediction to the input

where ***f*** is the model to explain

Temperature downscaling over North America

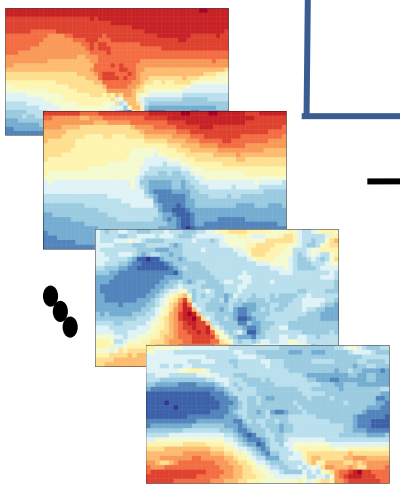
Deep learning models

CNN-DeepESD
CNN-PAN

UNet

Convolution + dense

Fully-convolutional

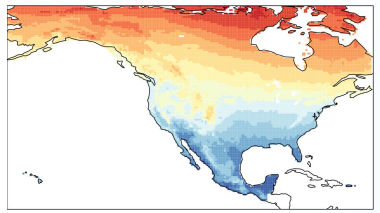


ERA-Interim (2° resolution)

5 thermodynamical variables
× 4 different vertical levels

Deep Learning

PP Empirical Statistical Downscaling



Observations (0.5° resolution)

Mean temperature (land)

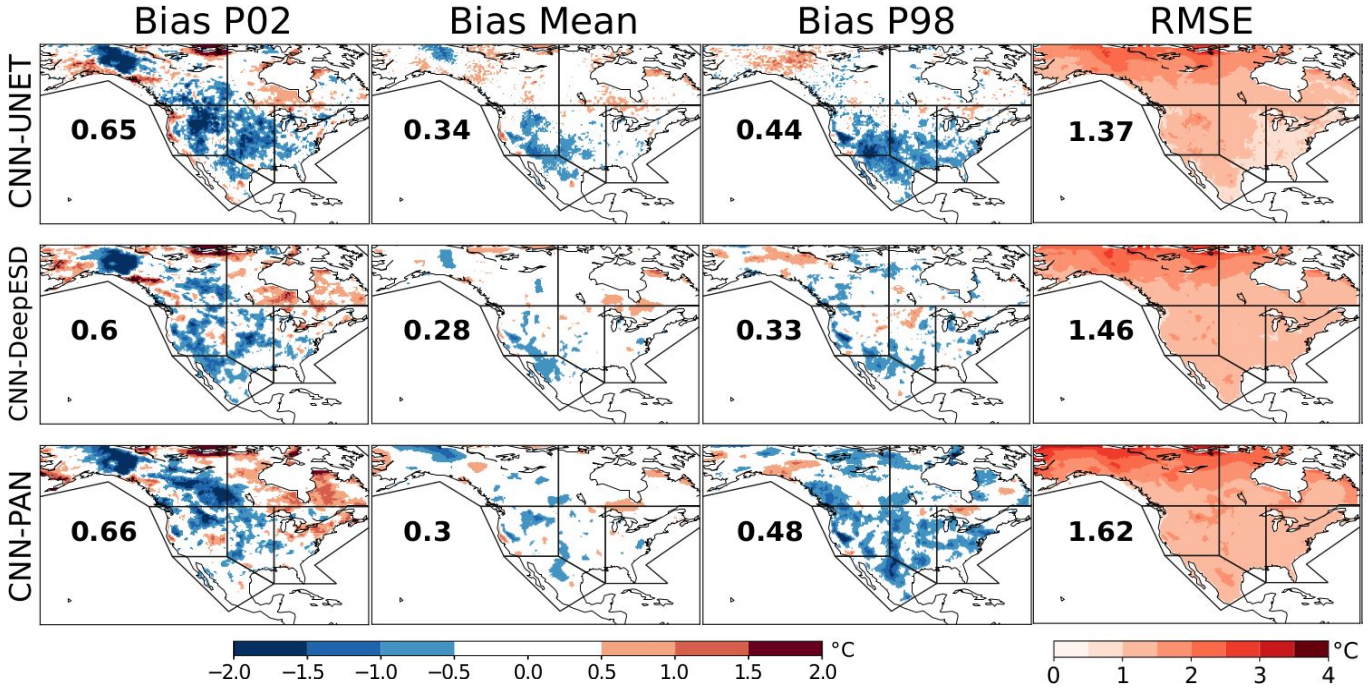
Training set
1979-2002

Test set
2003-2008

EC-Earth (CMIP5) RCP8.5

2006-2040
2041-2070
2071-2100

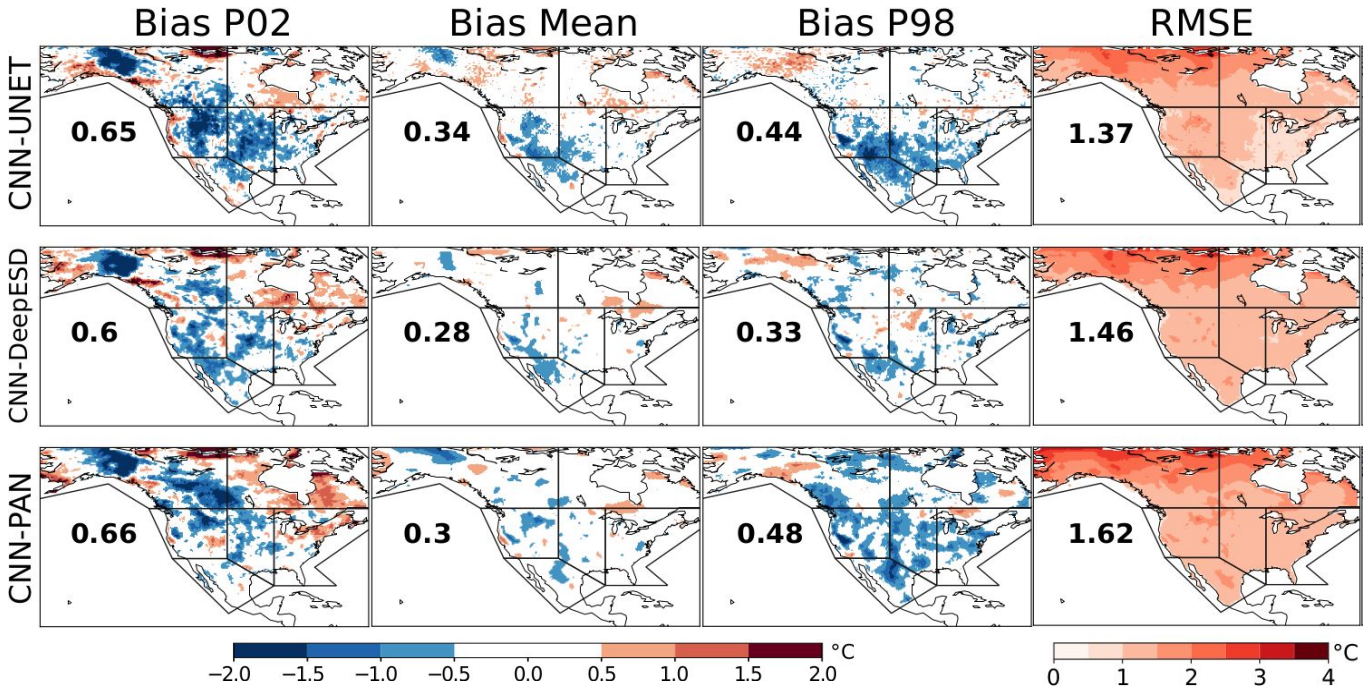
Evaluation on the test set

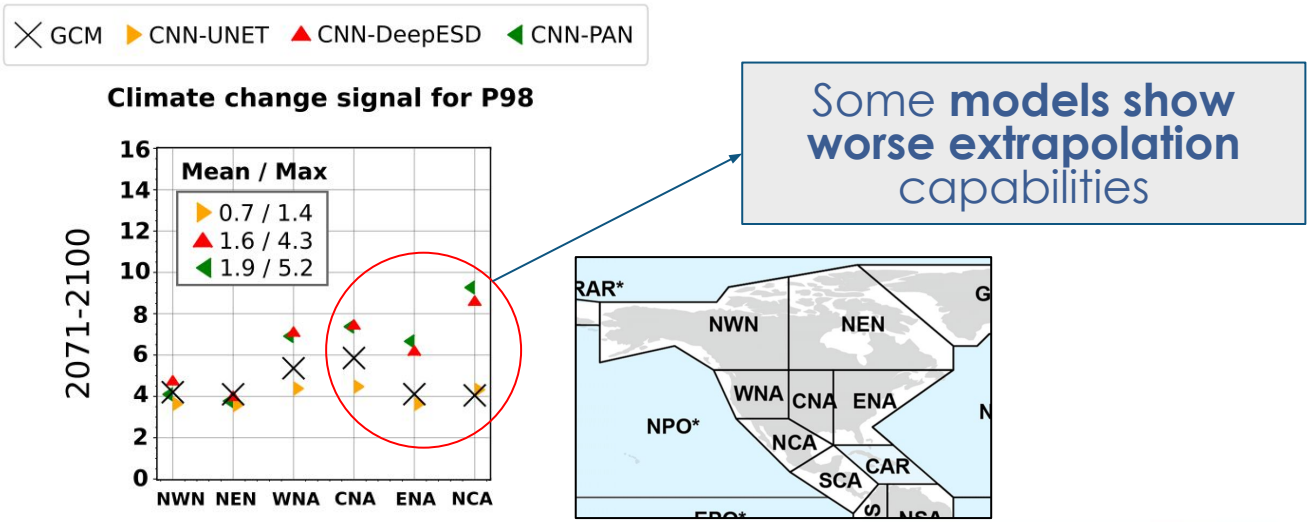


- **Similar spatial distribution** of biases across DL models

No best performing model for all regions and metrics

Evaluation on the test set





Some **models show worse extrapolation** capabilities

- **Similar spatial distribution** of biases across DL models

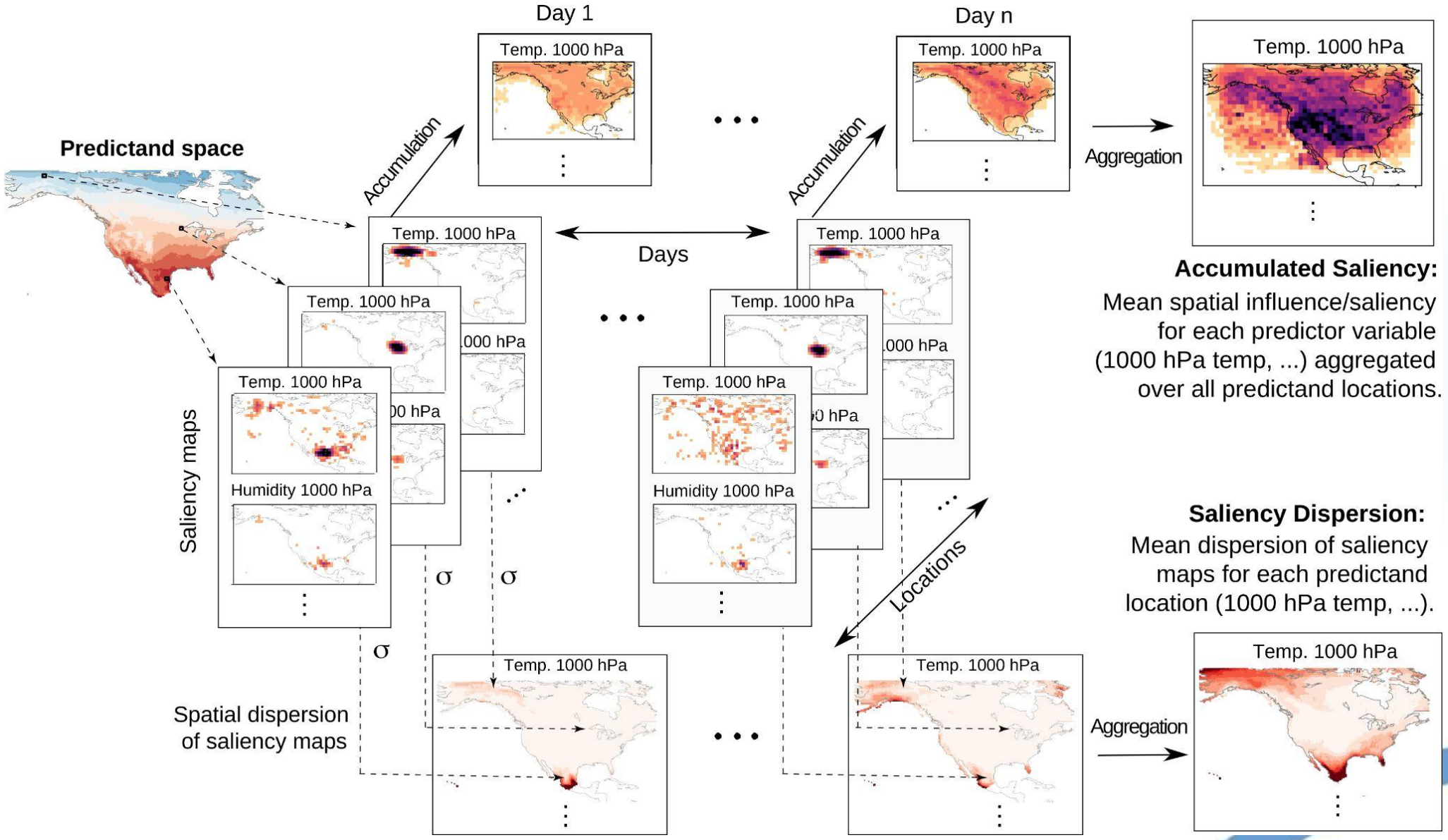
~~**No best performing model for all regions and metrics**~~

We need **further insights** on the **models behaviour**

↓

Interpretability

Interpretability-based metrics

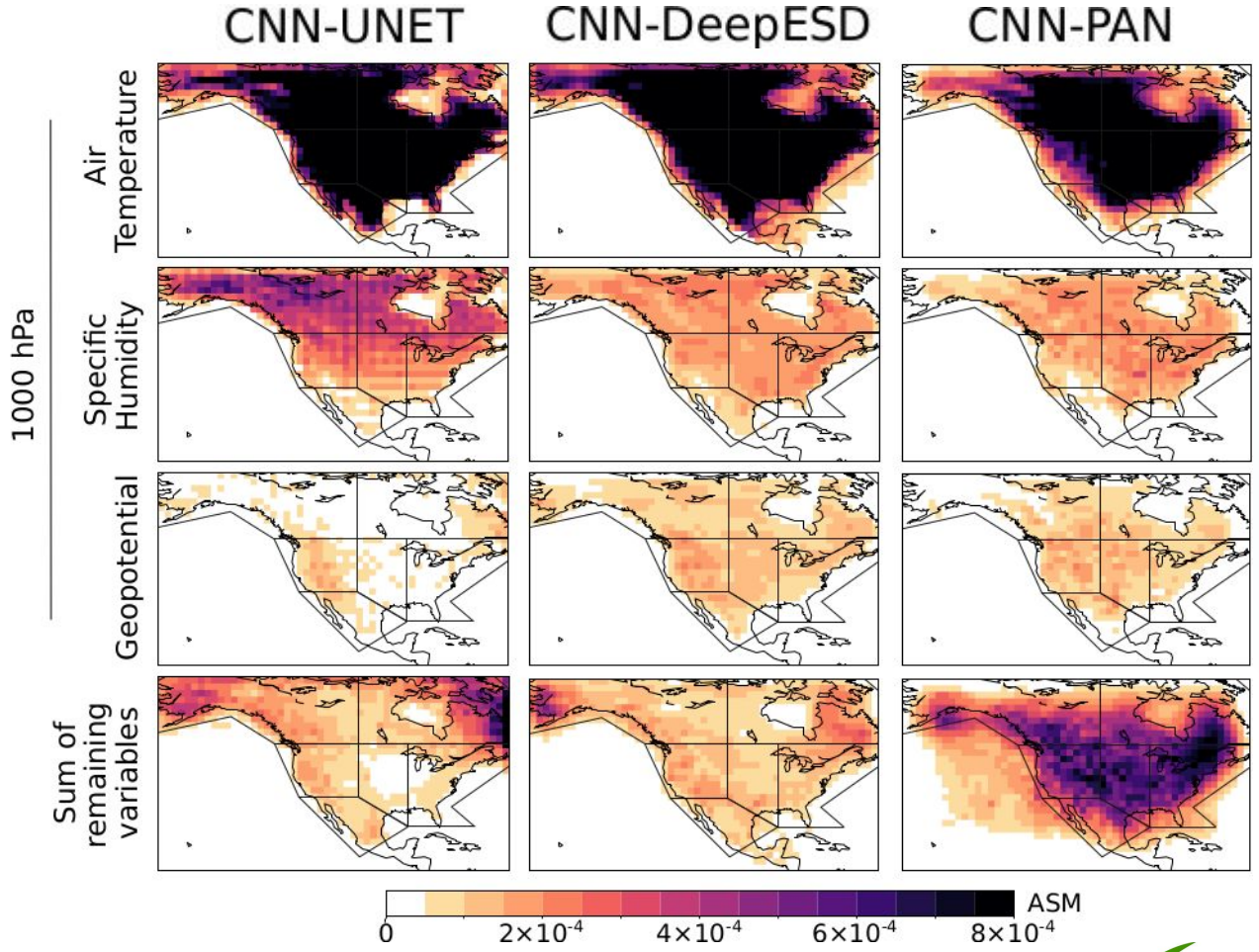


Accumulated Saliency:
Mean spatial influence/saliency for each predictor variable (1000 hPa temp, ...) aggregated over all predictand locations.

Saliency Dispersion:
Mean dispersion of saliency maps for each predictand location (1000 hPa temp, ...).

Source: González-Abad, J., Baño-Medina, J., & Gutiérrez, J. M. (2023). Using Explainability to Inform Statistical Downscaling Based on Deep Learning Beyond Standard Validation Approaches. arXiv preprint arXiv:2302.01771.

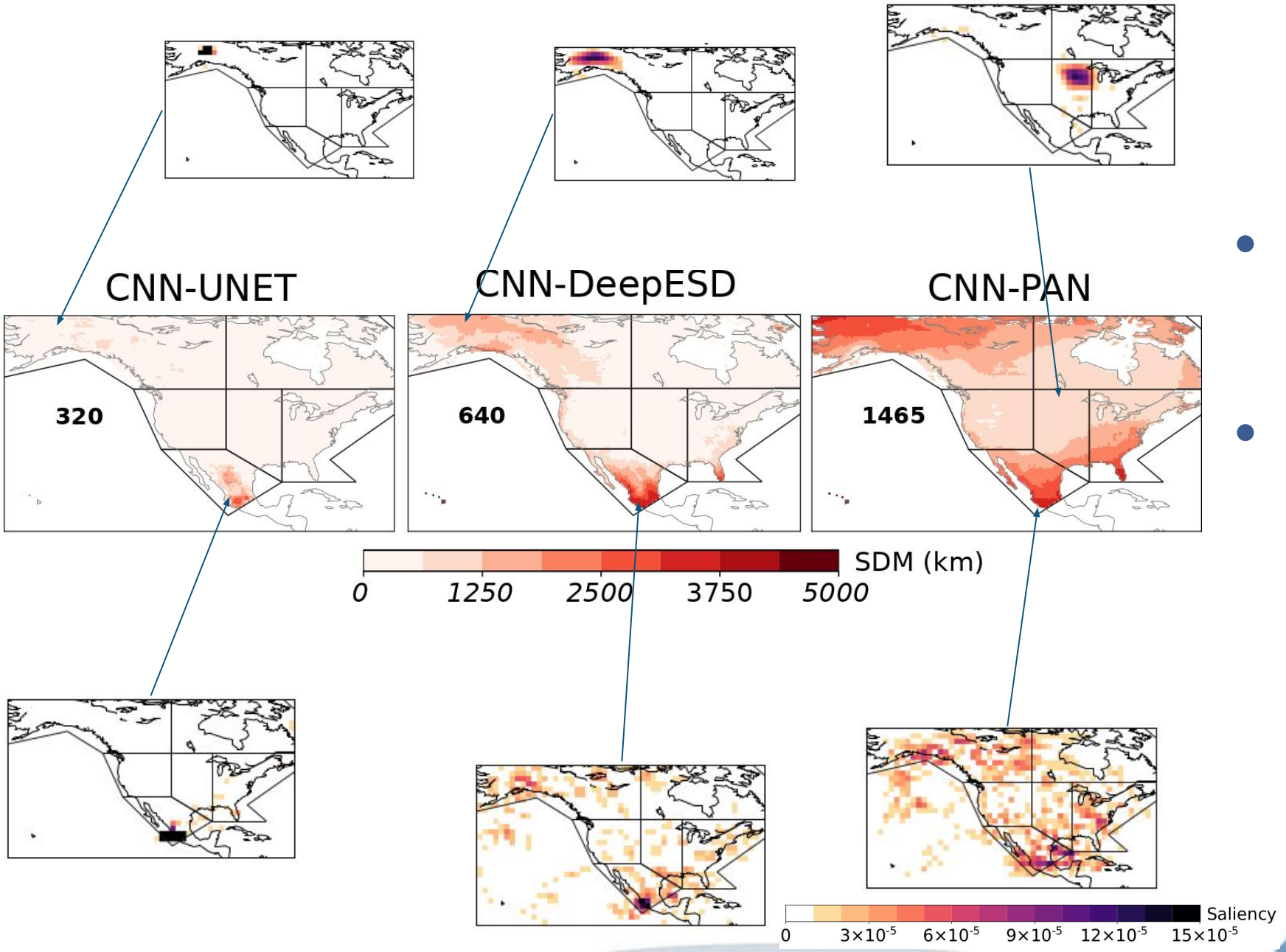
Accumulated saliency map (ASM)



- **Air temperature at 1000 hPa** is the **most relevant variable** → ✓ across seasons and DL models

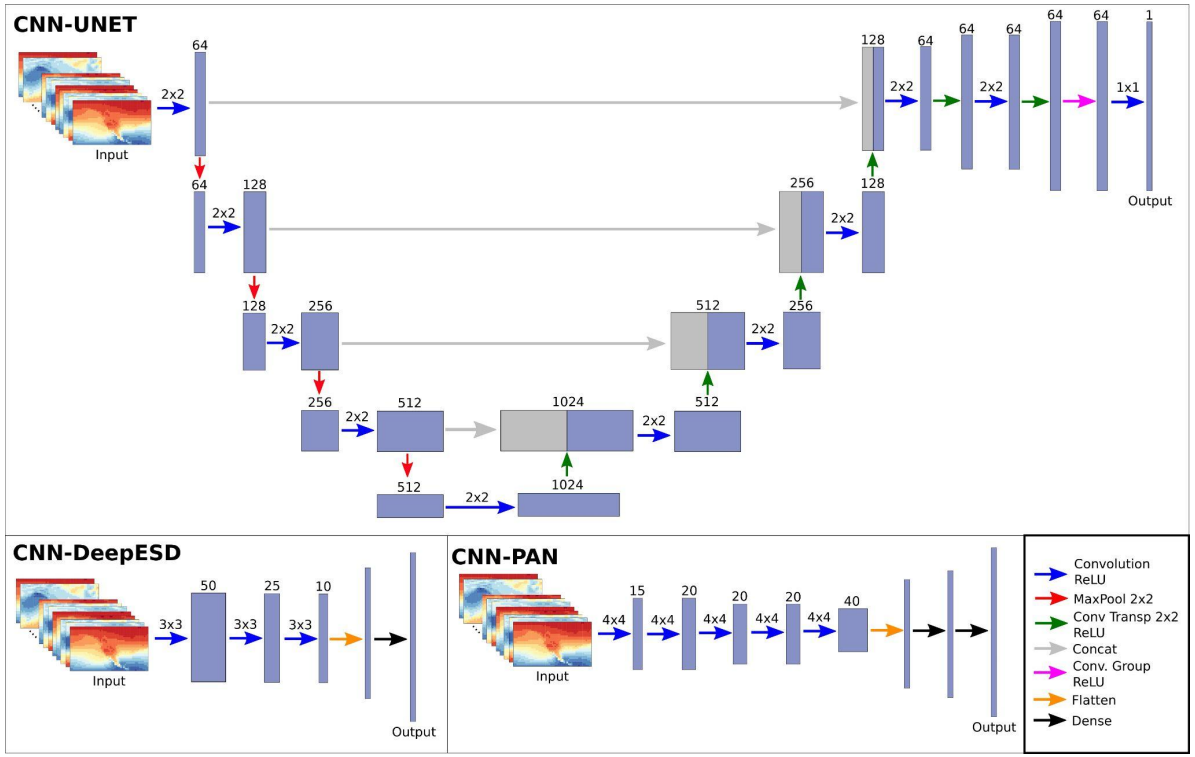
→ **It aligns with the scientific literature**

Saliency dispersion map (SDM)



- **CNN-UNET** shows **local patterns** across the region
- **CNN-DeepESD** and **CNN-PAN** shows **non-local patterns** for the **southern region** (NCA)

Overfitting of CNN-DeepESD and CNN-PAN



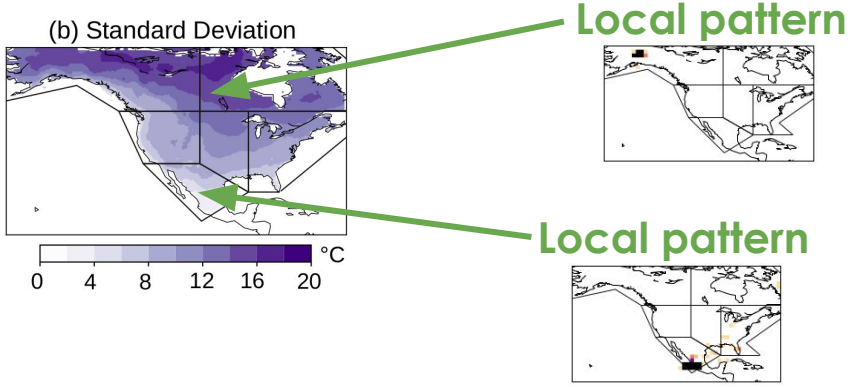
- **Convolutional layers** incorporate an inductive bias toward capturing **local relationships**
- **Dense layers** allow learning **non-local relationships**

CNN-UNET is implicitly local

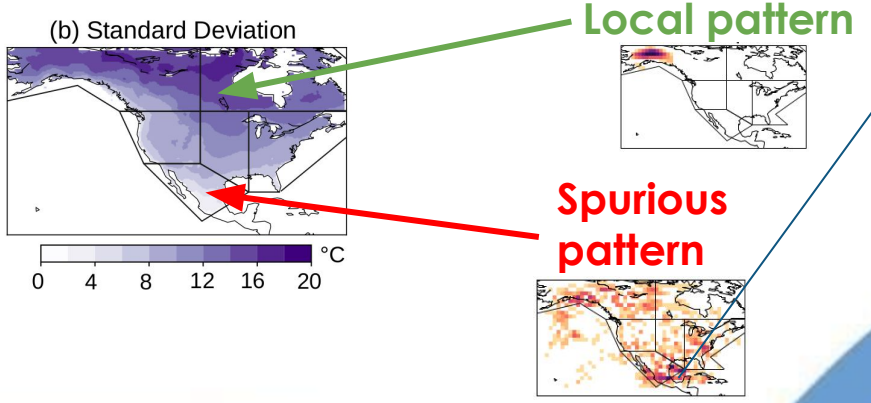
Temperature downscaling is local

Plausible relationships

CNN-UNET



CNN-DeepESD / CNN-PAN



Non-plausible relationships

Conclusions

- **Interpretability techniques** allow **expanding the standard evaluation techniques** for assessing deep learning models in statistical downscaling

- **Interpretability techniques** can help us gain **confidence** in deep learning models when **extrapolating** to GCMs in future scenarios

- Although **interpretability techniques** can be useful in offering transparency for DL models, they must be used with **caution** given their **limitations**

Can we **trust** these models?

→

Standard validation metrics
+
Interpretability

Instituto de Física de Cantabria

CSIC Consejo Superior de Investigaciones Científicas

UC Universidad de Cantabria

WCRP CORDEX Coordinated Regional Climate Downscaling Experiment

Thank you!

Jose González-Abad
gonzabad@ifca.unican.es

Jorge Baño-Medina
bmedina@ifca.unican.es

José Manuel Gutiérrez
gutierjm@ifca.unican.es