Unmasking Compensating Biases: A Process-Partitioned Neural Network Approach

Yuiko Ichikawa[1] and Miles Cranmer[2]

[1] University of Cambridge MPhil Data Intensive Science; University of Oxford

[2] University of Cambridge

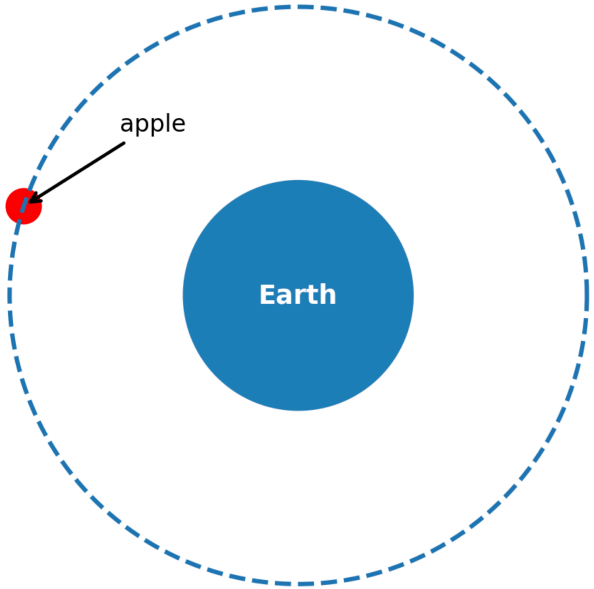
When we want to understand a system governed by physical laws, what do we do?

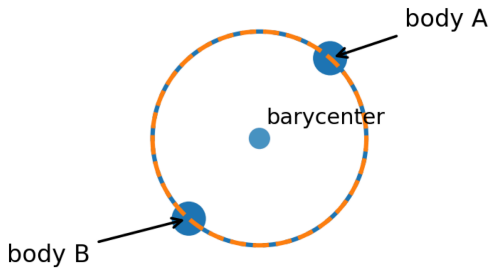
We start from observation.

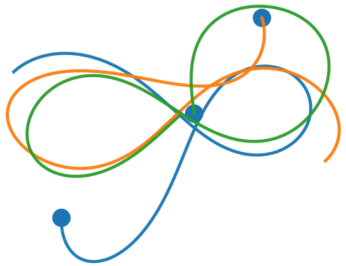
Problem

In many systems, multiple physical processes operate simultaneously.

That makes it difficult to disentangle their individual contributions.







Problem

In meteorology, many processes act at the same time.

Even when we model them separately, it is often unclear how much each process contributes to the final outcome.

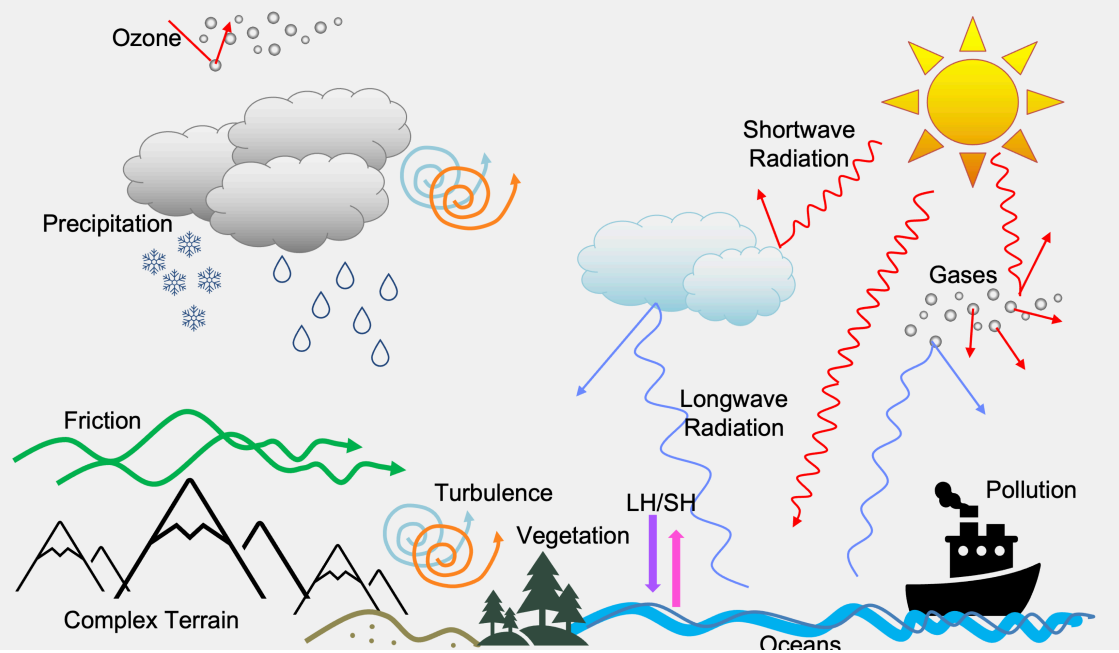
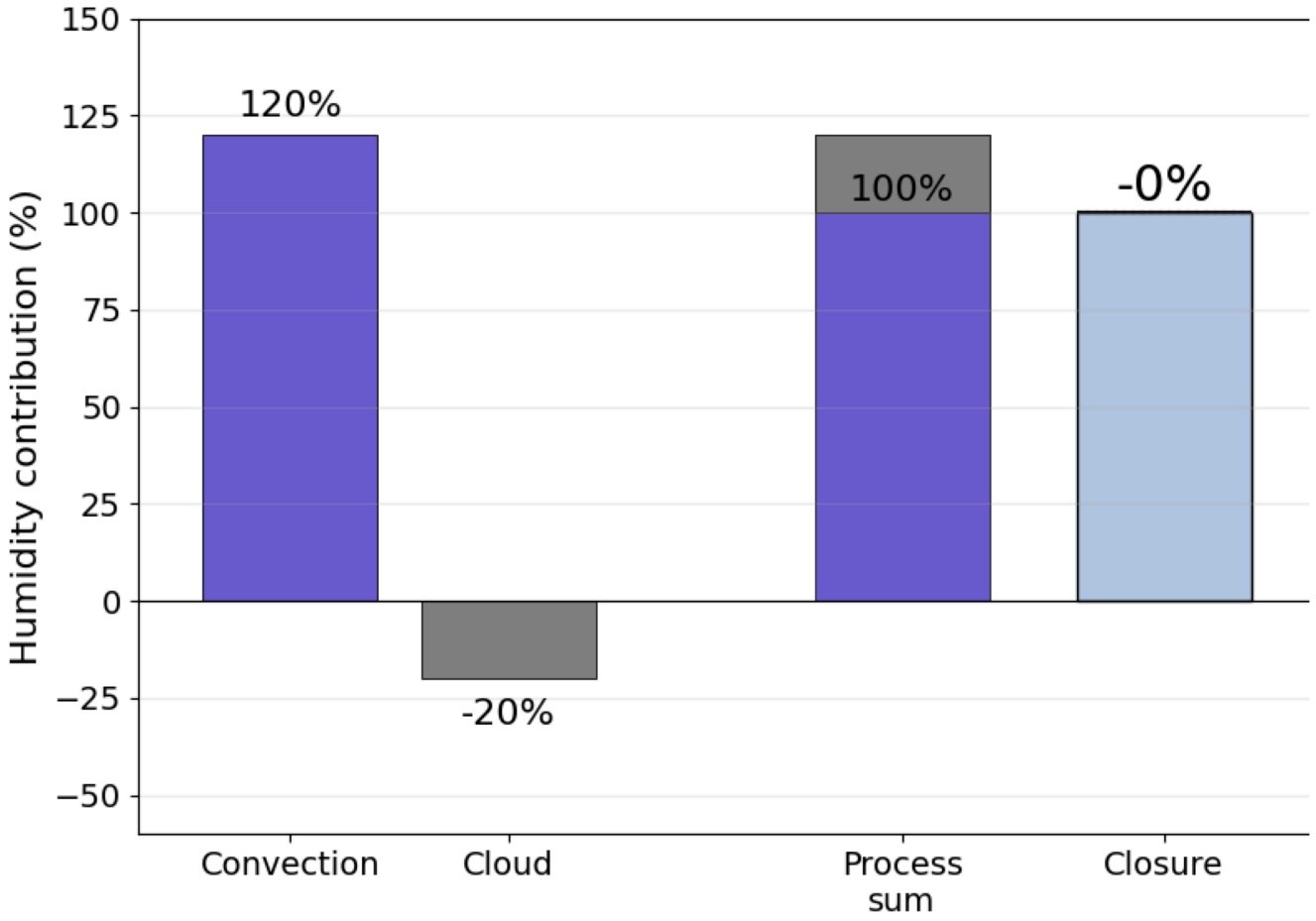


Figure source: https://www2.mmm.ucar.edu/wrf/users/wrf_users_guide/build/html/physics.html

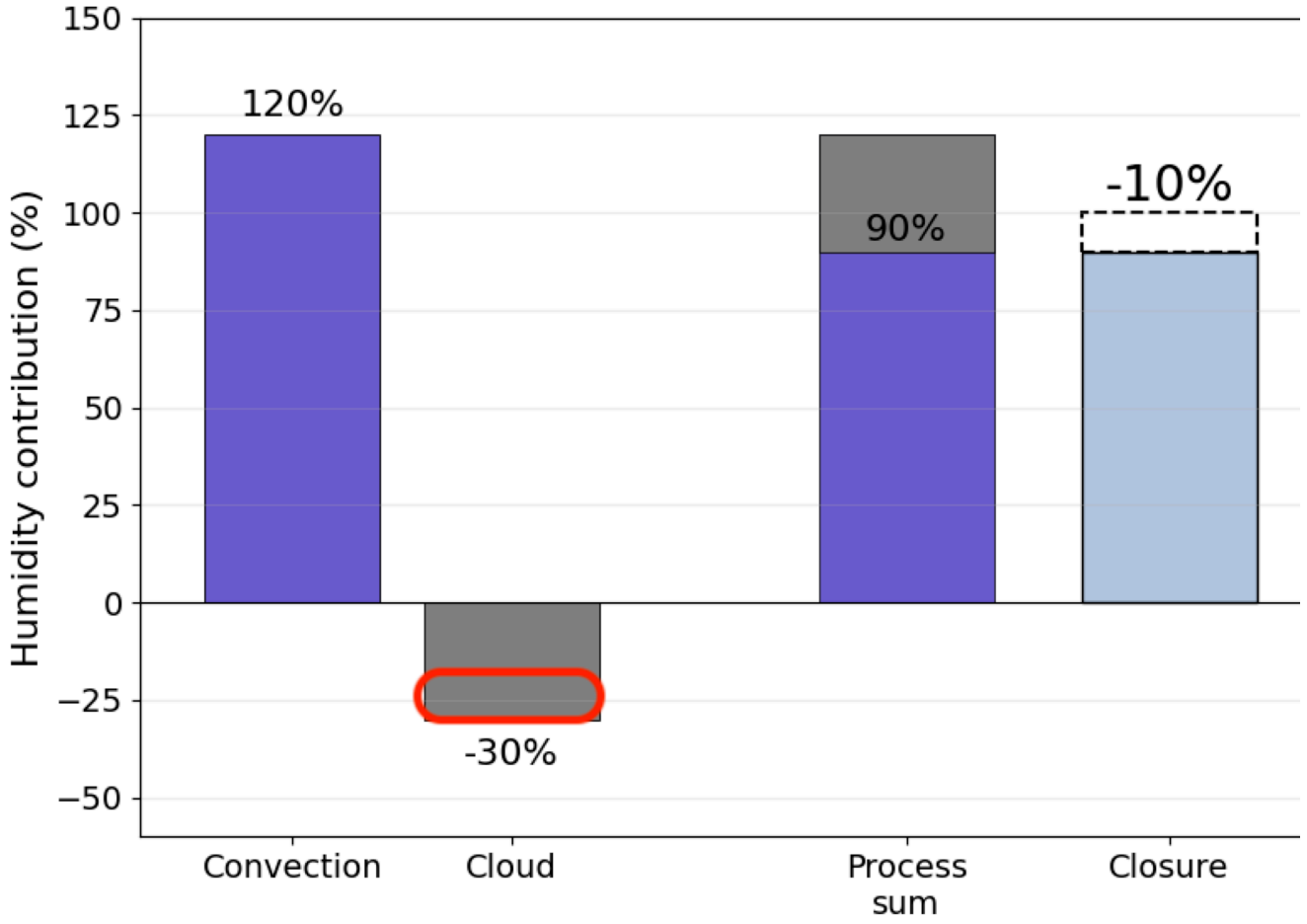
Ideal Solution

We decompose the total outcome into contributions from each process.



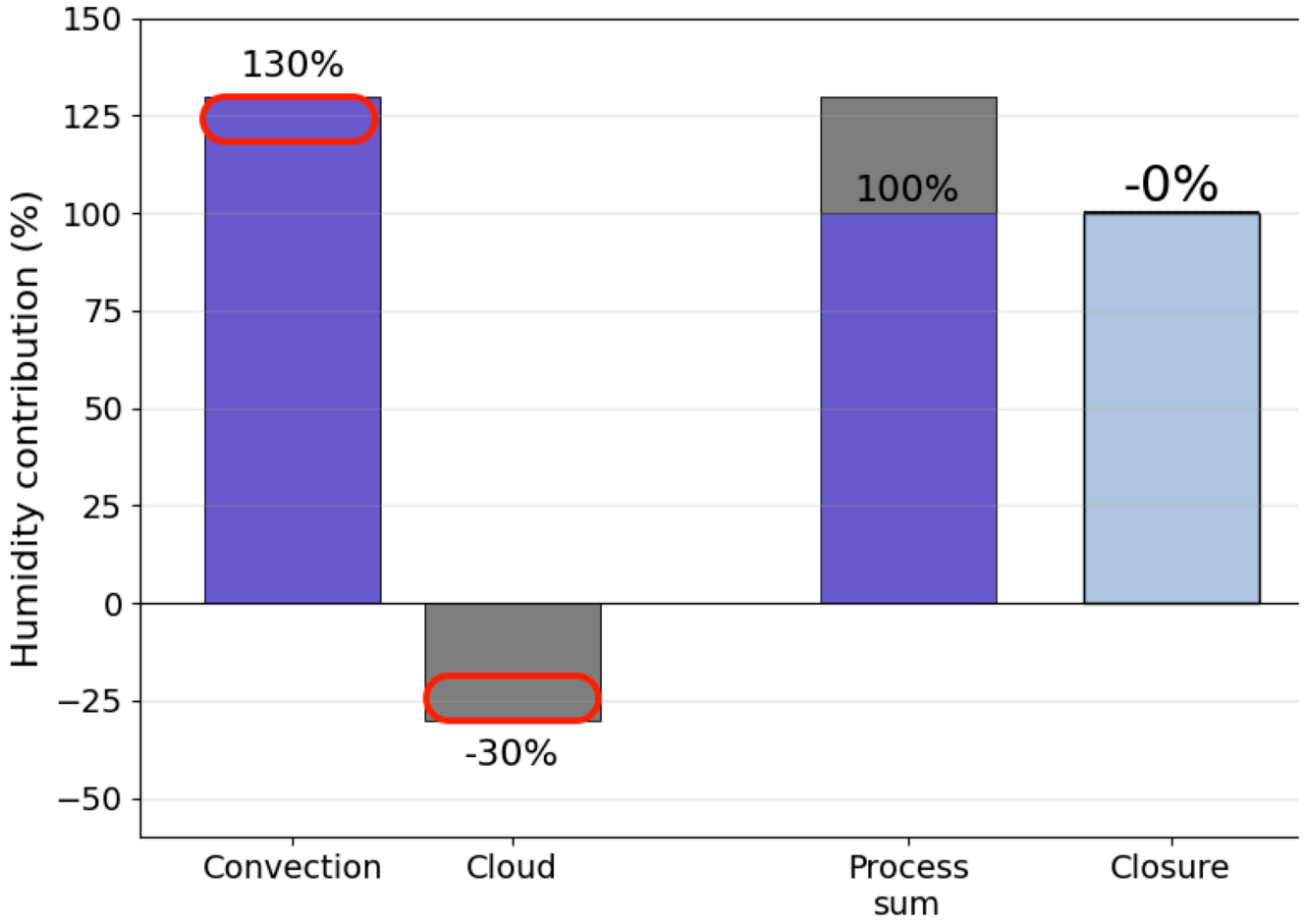
Practical Mismatch

In practice, this decomposition often does not match the observed outcome.



Compensation Error

To match observations, one process is often tuned to compensate for errors in another.



Example:

- Cloud tendencies are biased.
- Convection is tuned to compensate.
- Both cloud and convection climatologies become physically inconsistent, even if the total climatology looks correct.
- Apparent model skill can improve, but physical consistency degrades.

Later:

- Cloud physics is improved.
- Physical consistency improves locally, but overall bias re-emerges.
- Model performance can degrade.

This is compensation error.

See, e.g., Stevens et al. (2012) and Mauritsen et al. (2017).

Key Question

To avoid this problem, we need to quantify each process contribution.

But how can we do that?

Qualitative process-level understanding is often possible, but translating it into quantitative attribution remains difficult.

Previous Approaches

(1) Physics-based perturbation (PRP)

- Perturb one variable at a time in the model.
- Measure the resulting output change.
- Widely used in cloud and radiation studies (e.g., Wetherald and Manabe 1988).

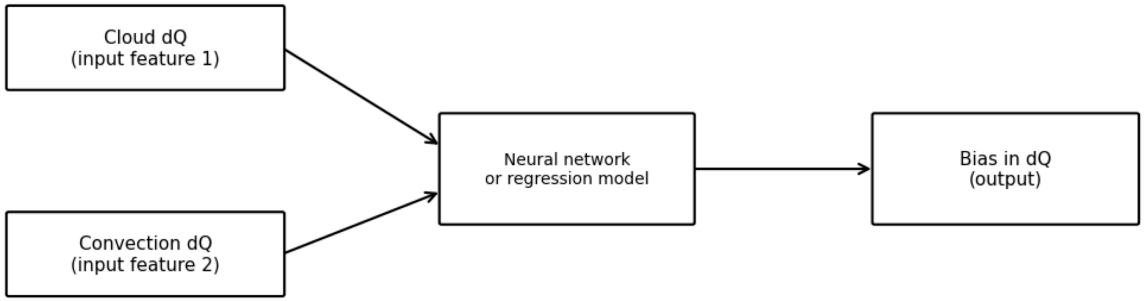
Limitations:

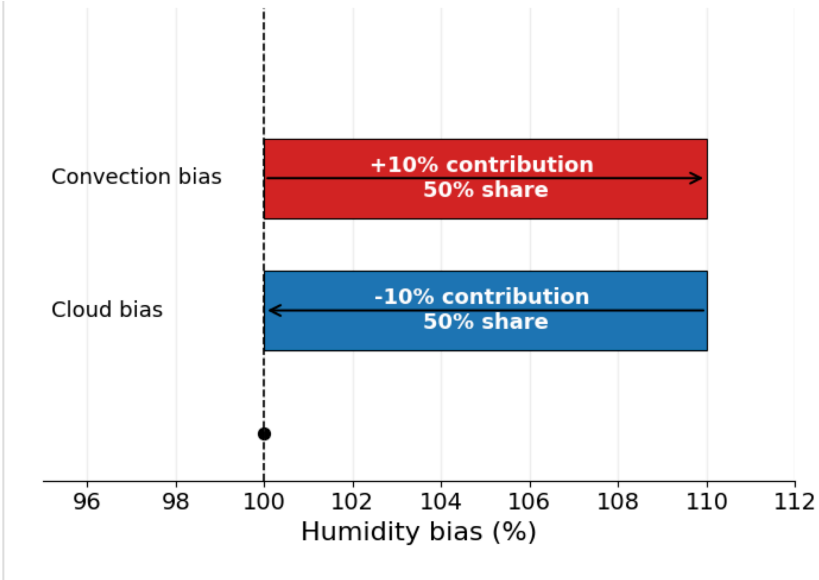
- Computationally expensive and labor-intensive for full NWP models.

Previous Approaches

(2) SHAP (Feature Attribution)

- Decomposes predictions into additive feature contributions.
- Widely used in explainable AI and increasingly in geoscience applications (e.g., Lundberg et al. 2020; Clare et al. 2022).

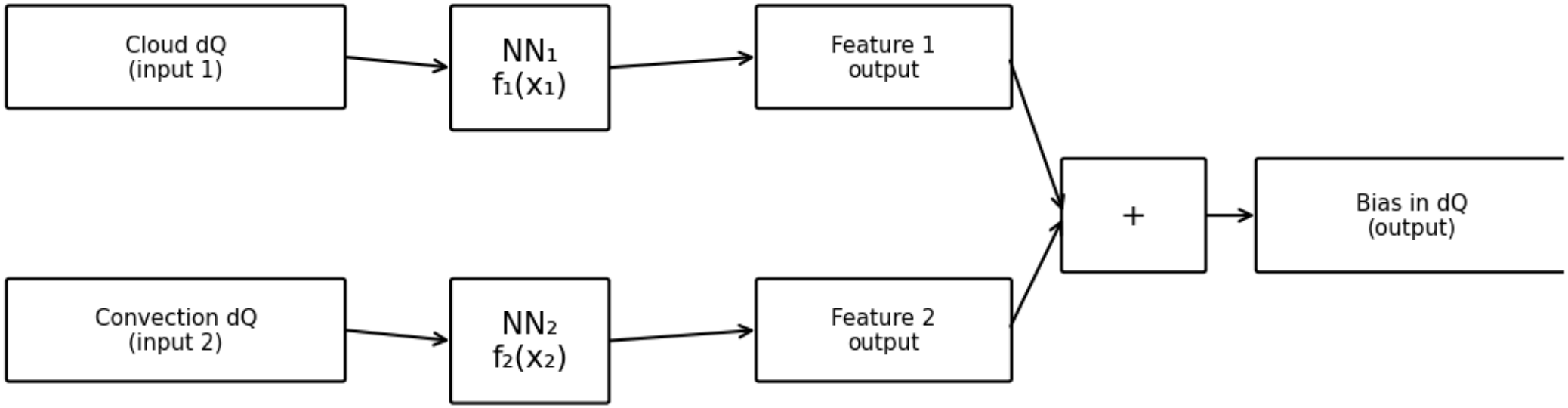




Limitations: Exponential increase of compute with the number of input features

Structural Motivation

Why should the network be process-partitioned?

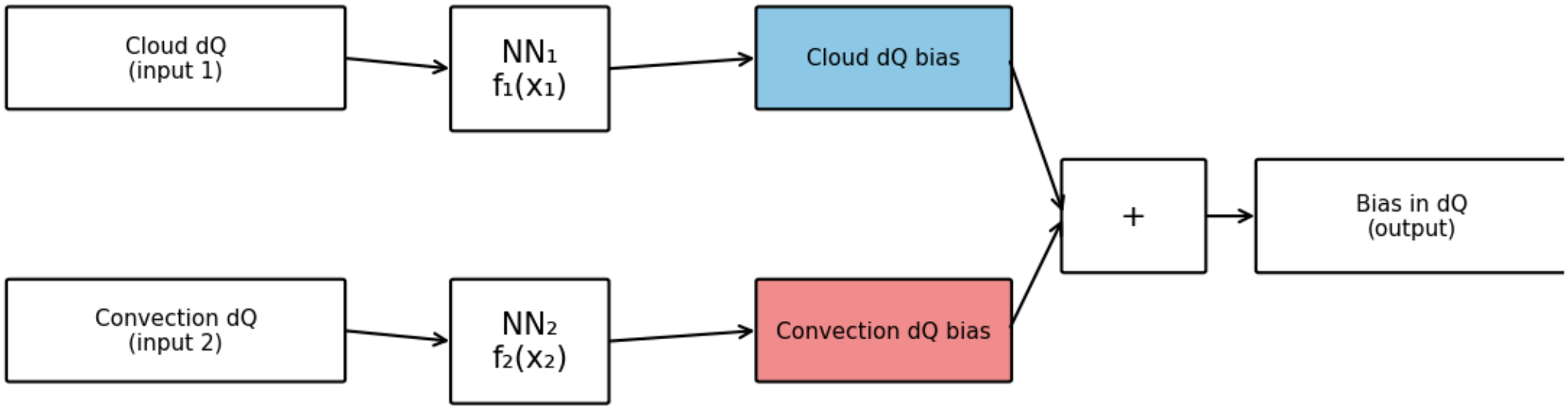


$$\text{Bias} = f_1(\text{Cloud dQ}) + f_2(\text{Convection dQ})$$

Within a single model step, most NWP models compute tendencies for each physical

Proposed Idea

Use a process-partitioned neural network (PPNN), where each branch corresponds to one physical process.



$$\text{Bias} = f_1(\text{Cloud dQ}) + f_2(\text{Convection dQ})$$

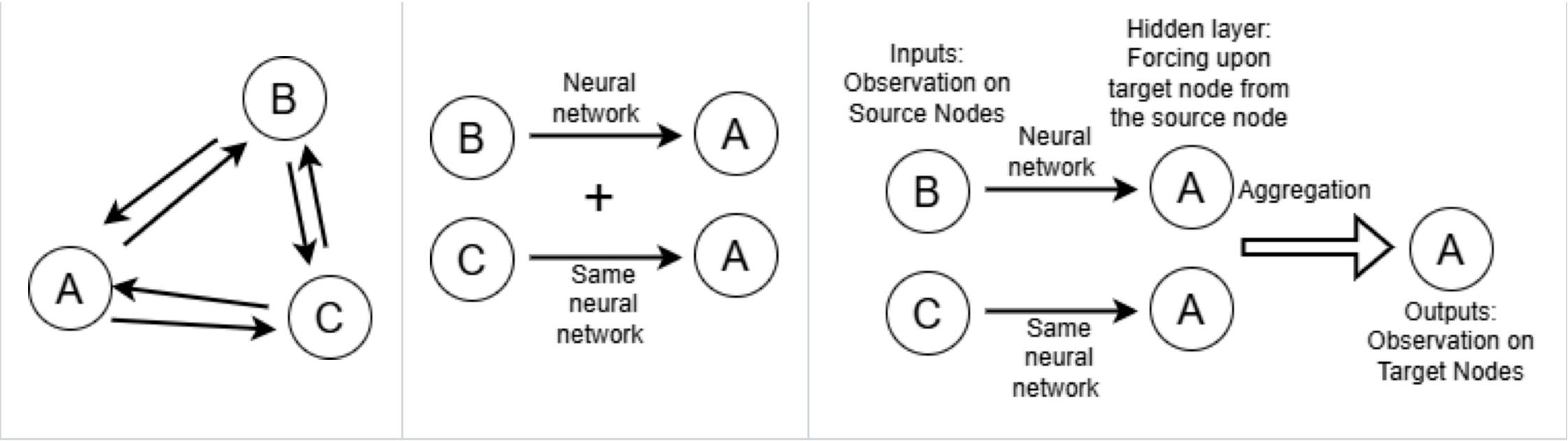
Why This Is Better

- Predicts bias (including case-by-case error), not only attribution summaries.
- Compatible with modern attribution tricks (What can be done for SHAP estimation can often be done on this network).
- Much cheaper than feature-wise attribution because decomposition is process-wise.

Did It Work?

Yes, in N-body learning problems (Cranmer et al. 2020).

- Input: positions at time `t`
- Target: positions at time `t+1`



Neural networks can learn these interactions for up to 16 bodies.

Test Setup

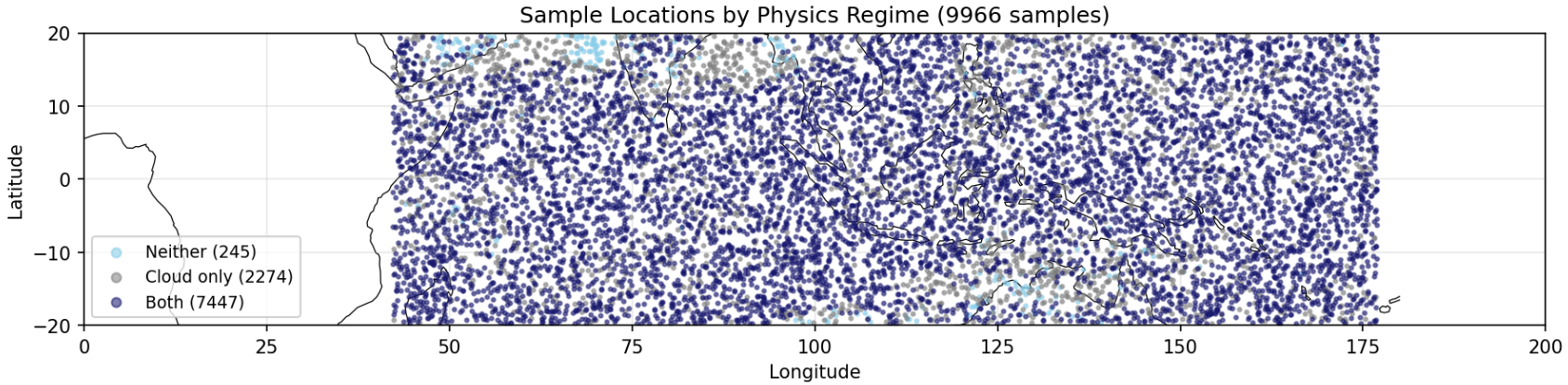
NWP Model Setting

OpenIFS single-column model (`cy48r1`) with 90 vertical levels.

Dataset

ML Input

- 10,000 tropical-ocean points sampled over 24 hours.
- Inputs: tendencies from two consecutive 4.5-minute steps (`Δt = 9 min`).



Why use a short time step?

- It isolates single-process impacts before long-term averaging.

- It captures rapid convective/cloud bias emergence in the first few steps.

Dataset

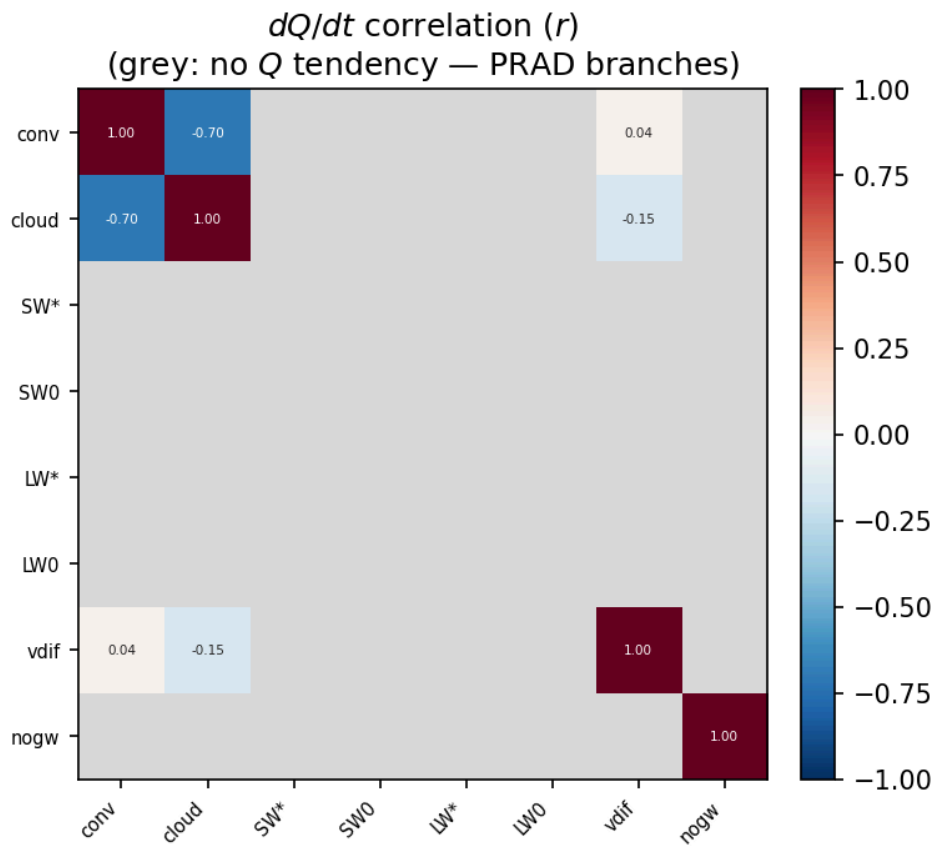
ML Target

We learn model bias as a tendency difference from a reference:

1. A synthetic target with a known relationship (controlled experiment).
2. The difference between OpenIFS and high-resolution Unified model output (Christensen et al. 2018).

Variables and Processes

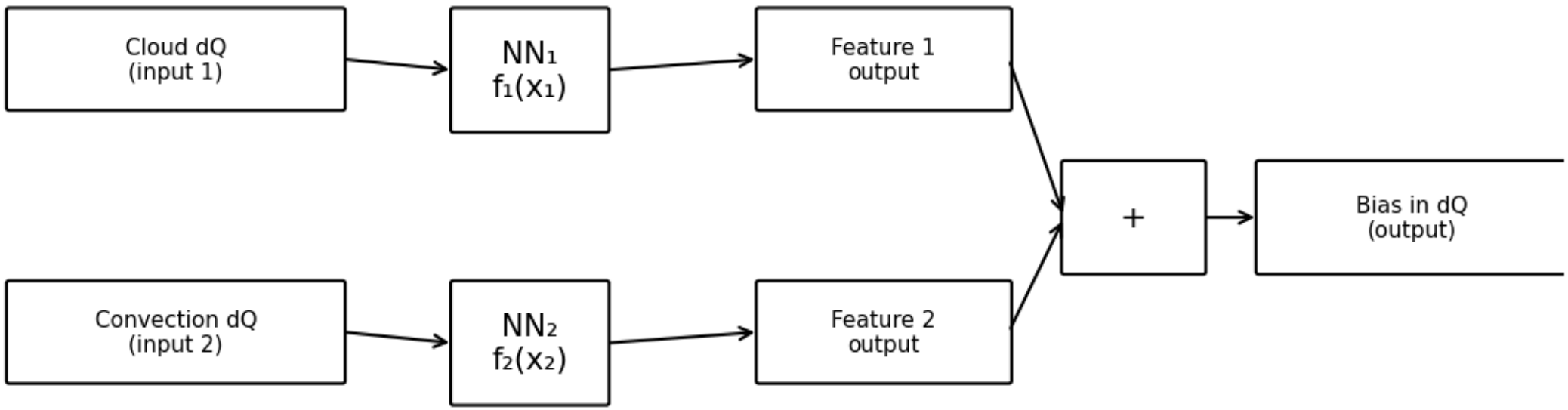
Included Processes	Tested Variable
Cumulus convection Cloud microphysics	Specific humidity (Q)



Simple enough to validate the concept, yet still challenging because of strong inter-process input correlation.

Model Architecture 1

Given the process-partitioned framework, what should each branch network look like?



$$\text{Bias} = f_1(\text{Cloud dQ}) + f_2(\text{Convection dQ})$$

Model Architecture 2

Forward Flow

Stage	What happens
Multi-scale encoder	Extract local and broader vertical patterns in parallel
Bottleneck	Compress information into a compact latent space
Decoder	Reconstruct branch signals from latent features
Head	Project to branch error tendencies

- Kept the architecture as simple as possible, with a small number of parameters.
- Suggestions for alternative structures are welcome.
- Dropout ratio: 20%.

Model Architecture 3

We use an informal commonality-analysis approach (Seibold and McPhee 1979) to better attribute signal across correlated features.

- Calculate the sample variance of each predicted branch error.
- Use the variance as contribution estimate; which works as a weighting term in the final output.
- Update the contribution estimate every 5 epochs after a 10-epoch warm-up.

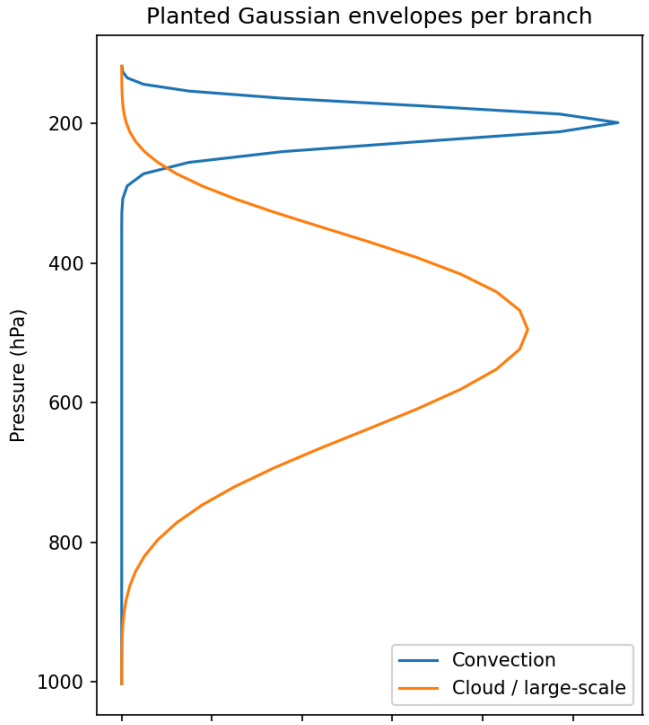
Model Training and Bias Evaluation

1. Train error prediction with 4-fold cross-validation for up to 40 epochs.
2. Analyze hidden representations, expected to correspond to process-level error/bias structure.

Result 1: Synthetic Bias Test

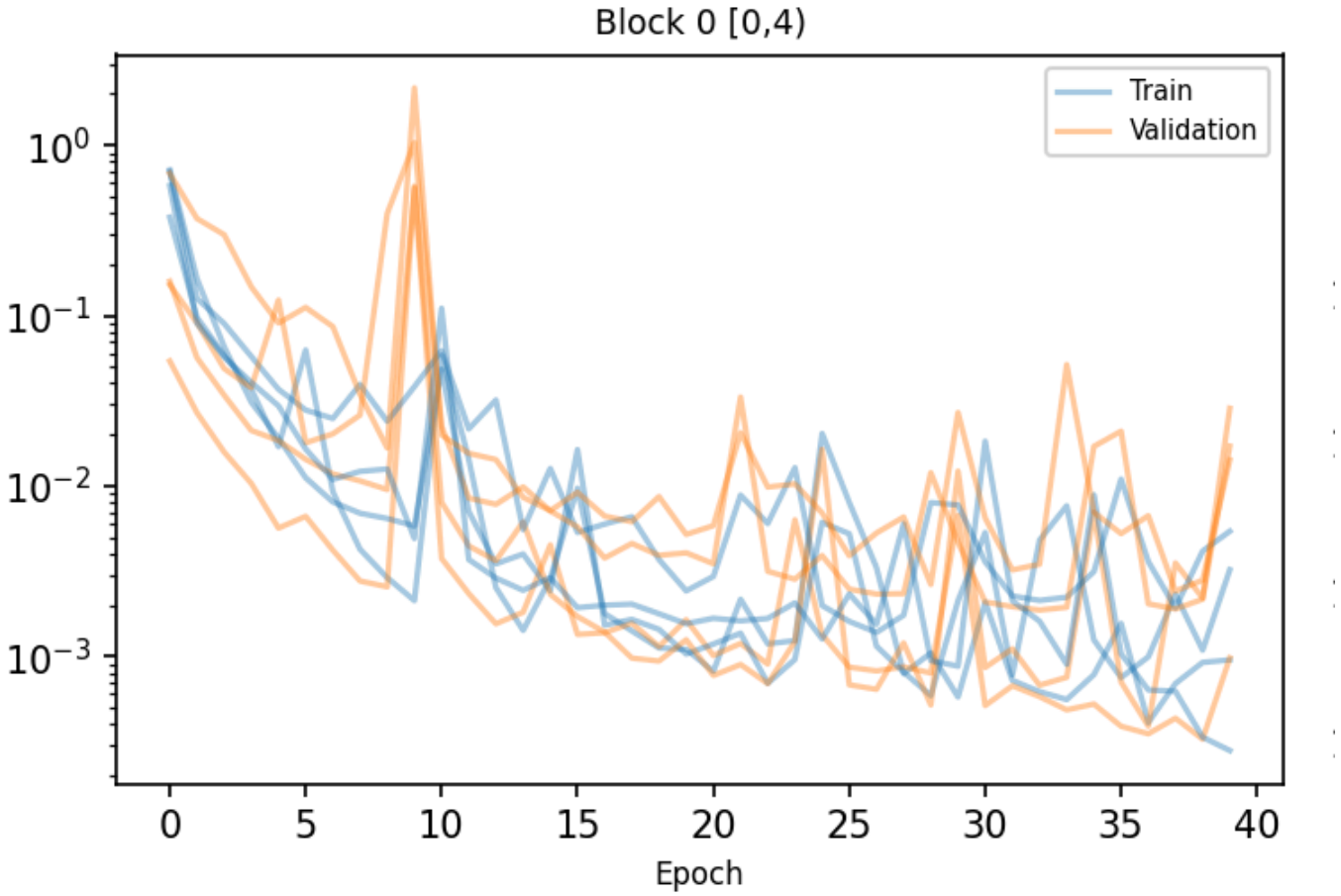
- Scenario: `dQ` from convection is biased near the tropopause; `dQ` from cloud is biased in the mid-troposphere.
- Prediction target:

```
dQ_bias = dQ_conv * conv_weight + dQ_cloud * cloud_weight
```



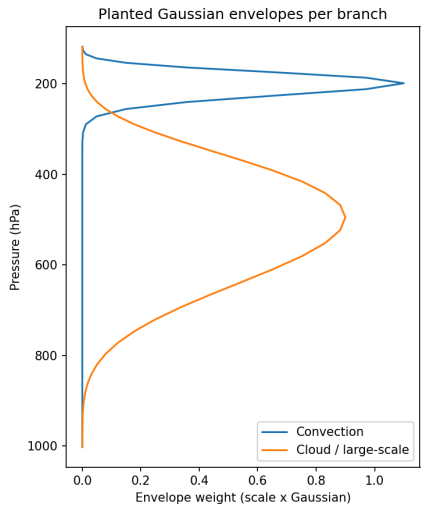
Result 1 (cont.): Synthetic Bias Test

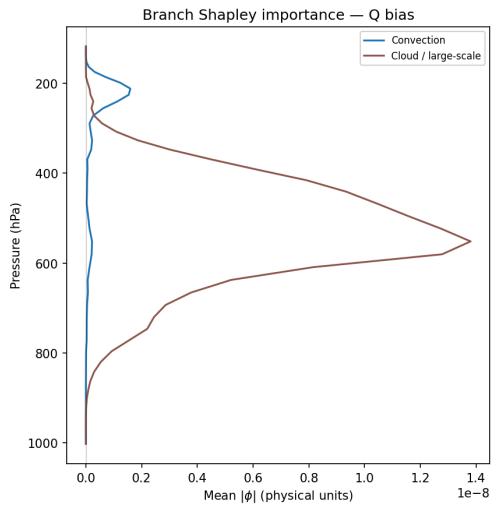
- Scenario: `dQ` from convection is biased near the tropopause; `dQ` from cloud is biased in the mid-troposphere.

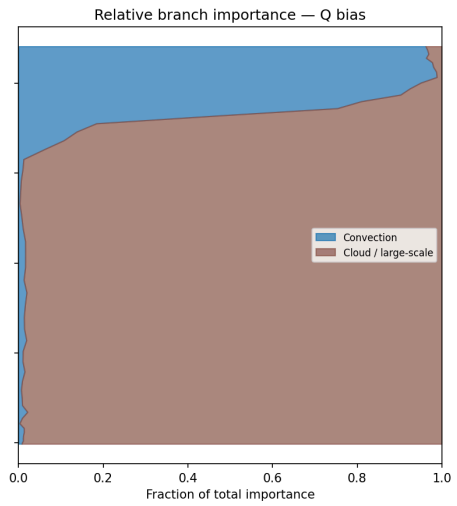


Result 1 (cont.): Synthetic Bias Test

Weight profiles and predicted contribution attribution generally agree.

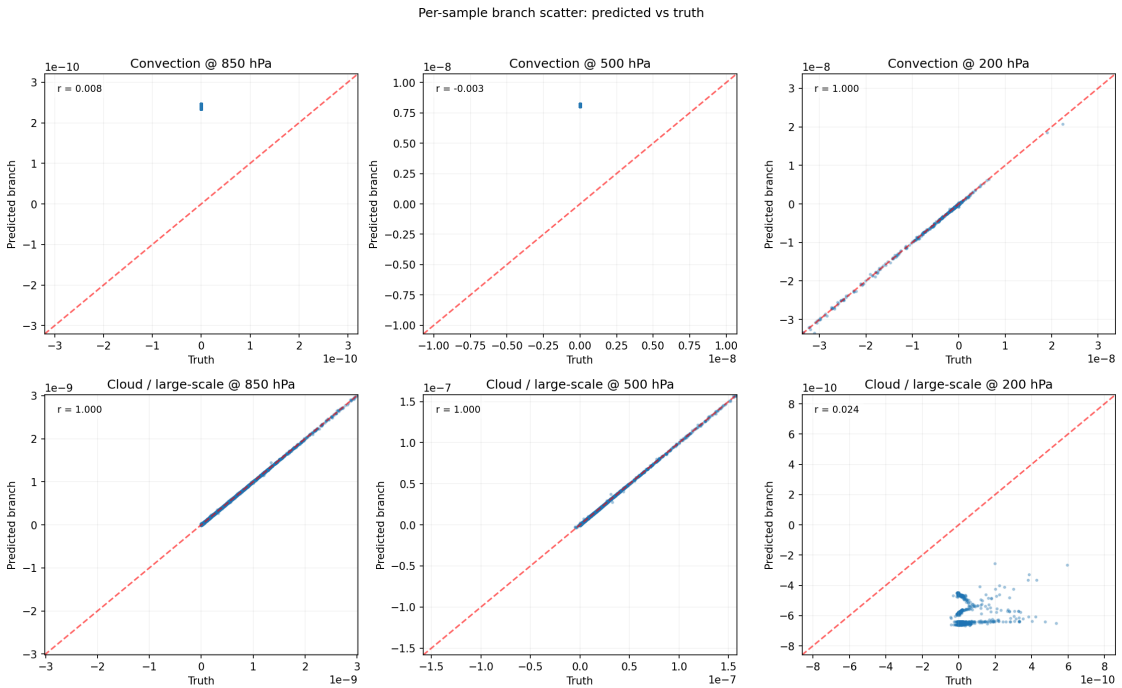






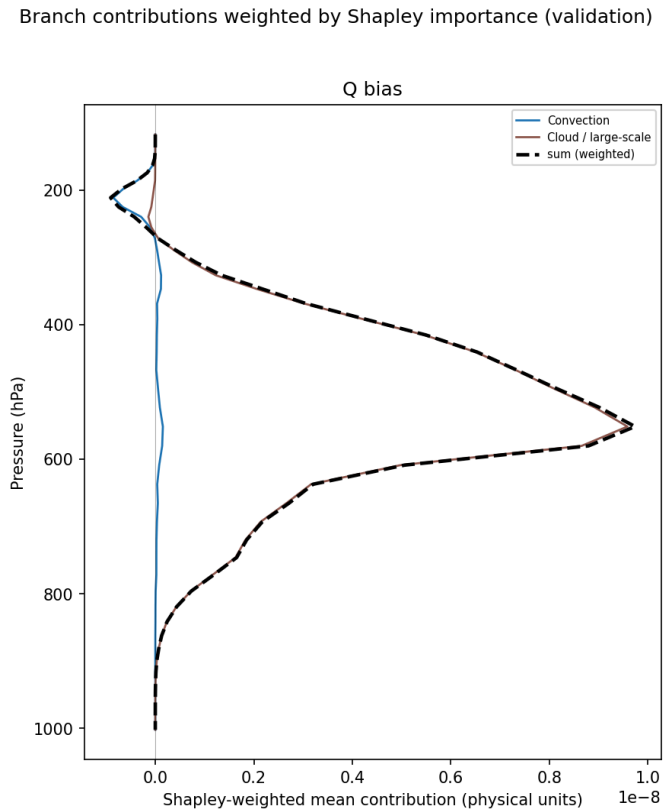
Result 1 (cont.): Synthetic Bias Test

- Scenario: `dQ` from convection is biased near the tropopause; `dQ` from cloud is biased in the mid-troposphere.
- It also predicts case-by-case error very well.



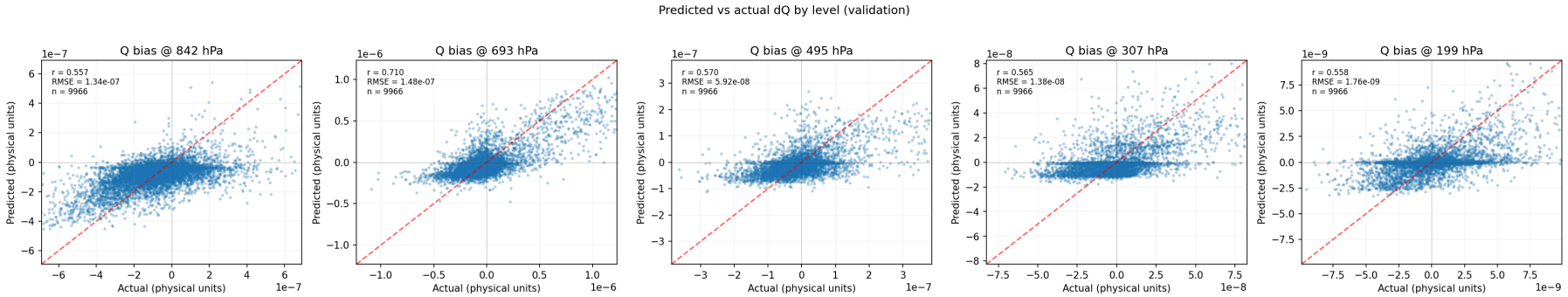
Result 1 (cont.): Synthetic Bias Test

- The model separates bias contributions from different processes.
- Performance can approach near-perfect separation.



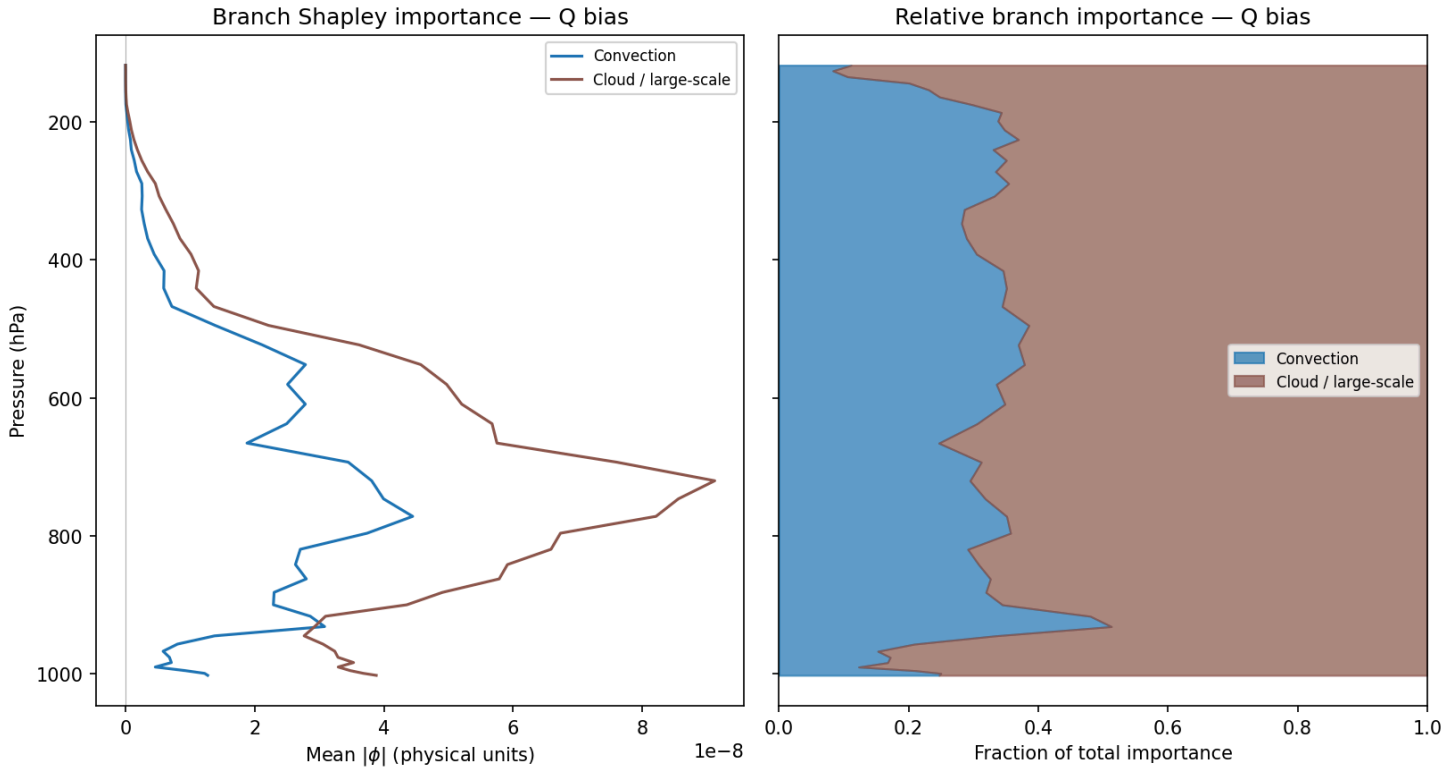
Result 2: Convection Permitting Model Test

- Scenario: Real-case test where the true process-level bias decomposition is unknown.



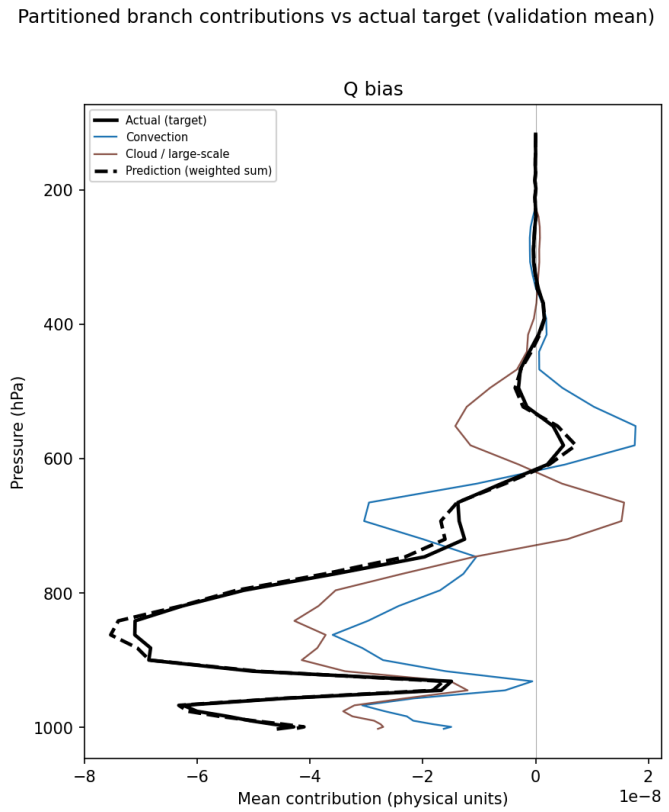
- Performance is not perfect, but predictions still agree reasonably well with observed bias (correlation coefficient around 0.6).

Result 2 (cont.): Convection Permitting Model Test



- The attribution is more likely to assign bias to cloud processes than to convection.

Result 2 (cont.): Convection Permitting Model Test



- Bias compensation appears in the upper troposphere.
- In the mid-troposphere, both convection and cloud branches show negative bias.

Summary

- Process-partitioned neural networks provide an efficient, interpretable way to isolate process contributions.
- Tested on an OpenIFS-based SCM.
- Successfully attributed synthetic bias.
- Tests against convection-permitting output indicate compensation in the upper troposphere.

Possible next steps: using it against field campaigns

If the underlying bias structure is as simple as in the synthetic tests, the method is already suitable for single-point, month-long field-campaign datasets (roughly 10,000 samples -- e.g. TWPICE).

- Performance with 10,000 samples suggests that scaling to intensive field-campaign datasets is feasible.
- The same framework could also be used to diagnose region-specific bias characteristics.

Possible next steps: using it against larger sample + stronger model

If the real bias structure is more complex, the next step is to strengthen both the attribution method and the training setup.

- Test branch dropout; Game theory based branch weighting; other attribution tricks (suggestions welcome).
- Increase the training dataset.
- Increase model capacity and complexity (convolution network with more channel, try Transformer variants).

Link to the github branch with current model code

https://github.com/iyui1223/pann-training

References

Clare, Mariana C. A., Maike Sonnewald, Redouane Lguensat, Julie Deshayes, and V. Balaji. 2022. "Explainable Artificial Intelligence for Bayesian Neural Networks: Toward Trustworthy Predictions of Ocean Dynamics." *Journal of Advances in Modeling Earth Systems* 14 (11). https://doi.org/10.1029/2022MS003162.

Christensen, Hannah M., Andrew Dawson, and Christopher E. Holloway. 2018. "Forcing Single-Column Models Using High-Resolution Model Simulations." *Journal of Advances in Modeling Earth Systems* 10 (8): 1833-1857. https://doi.org/10.1029/2017MS001189.

Cranmer, Miles D., Alvaro Sanchez-Gonzalez, Peter W. Battaglia, Rui Xu, Kyle Cranmer, David N. Spergel, and Shirley Ho. 2020. "Discovering Symbolic Models from Deep Learning with Inductive Biases." *Advances in Neural Information Processing Systems* 33 (NeurIPS 2020). https://arxiv.org/abs/2006.11287.

References (cont.)

Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. "From Local Explanations to Global Understanding with Explainable AI for Trees." *Nature Machine Intelligence* 2 (1): 56-67. https://doi.org/10.1038/s42256-019-0138-9.

Mauritsen, Thorsten, Daniel Klocke, Lorenzo Tomassini, Frederic Hourdin, Andrew Gettelman, Jean-Christophe Golaz, Venkatramani Balaji, et al. 2017. "The Art and Science of Climate Model Tuning." *Bulletin of the American Meteorological Society* 98 (3): 589-602. https://doi.org/10.1175/BAMS-D-15-00135.1.

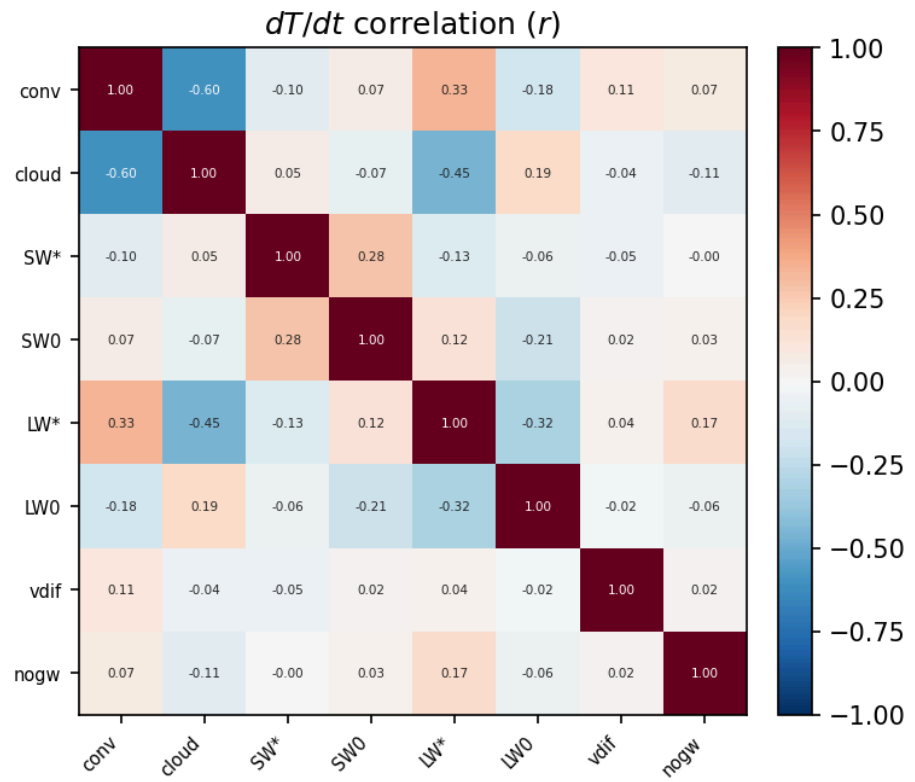
Seibold, David R., and Robert D. McPhee. 1979. "Commonality Analysis: A Method for Decomposing Explained Variance in Multiple Regression Analyses." *Human Communication Research* 5 (4): 355-365. https://doi.org/10.1111/j.1468-2958.1979.tb00649.x.

References (cont.)

Stevens, Bjorn, Erich Roeckner, Thorsten Mauritsen, Traute Crueger, Monika Esch, Marco Giorgetta, Helmuth Haak, et al. 2012. "Tuning the Climate of a Global Model." *Journal of Advances in Modeling Earth Systems* 4 (3). https://doi.org/10.1029/2012MS000154.

Wetherald, R. T., and S. Manabe. 1988. "Cloud Feedback Processes in a General Circulation Model." *Journal of the Atmospheric Sciences* 45 (8): 1397-1416. https://doi.org/10.1175/1520-0469(1988)045<1397:CFPIAG>2.0.CO;2.

Estimating dT is more challenging because more processes are involved



This can be tackled in a divide-and-conquer manner:

1. Estimate `dQ` bias for cloud and convection.
2. Fix `dQ`-bias learning parameters.
3. Estimate `dT` bias for clear-air columns.
4. Fix `dT`-bias learning parameters for clear-air conditions.
5. Estimate `dT` bias for cloud, convection, and SW/LW cloud-radiative effects, leaving only two fully unknown targets.