

Neural Network Architecture and Training

S1 CM1 simulations

Idealized TC simulations were conducted with the axisymmetric configuration of Cloud Model 1 (CM1, v21.1; Bryan and Fritsch, 2002), a non-hydrostatic, nonlinear model designed for idealized atmospheric-flow simulations. The model used 224 radial grid points and 72 vertical levels, with a radial domain extending to approximately 1316 km. The inner radial grid spacing was 2.5 km over the innermost 280 km and was gradually stretched to 16 km toward the outer boundary, aiming to resolve the TC using similar scales as the WRF simulations. In the vertical, the domain extended to 25 km, with grid spacing increasing from 20 m near the surface to 640 m aloft. All simulations used identical initial conditions from the default CM1 axisymmetric hurricane configuration (based on Bryan, 2012), on an f-plane with Coriolis parameter $f=5\times 10^{-5} \text{ s}^{-1}$. Moist processes were enabled using the Morrison double-moment microphysics scheme, without cumulus parameterization. Radiation and the PBL used the NASA-Goddard parametrization and YSU parametrization, respectively. Surface fluxes were computed over an all-ocean lower boundary using a fixed sea-surface/skin temperature of 301.15 K. The model was integrated for 8 days, with a time step of 4 s where model outputs were produced hourly.

Based on this reference configuration, a large ensemble of idealized axisymmetric simulations was designed to generate a broad range of atmospheric states. The purpose of these simulations was to sample a wide diversity of physically possible PV-structures along with their relevant wind fields, and potential temperature distributions. To this end, all possible combinations of multiplicative coefficients 0, 1, 2, 3 and 4 were applied at every model time step to selected parametrized temperature tendencies, namely (1) the temperature tendency due to microphysics, (2) the longwave-radiative temperature tendency, and (3) the planetary-boundary-layer temperature tendency. Coefficients were constant throughout the duration of simulations and each simulations was using one unique combination of multiplicative coefficients. A coefficient of 0 suppresses the corresponding tendency, a coefficient of 1 retains the default CM1 formulation, and coefficients larger than 1 amplify the contribution of that process.

To further enlarge the diversity of simulated atmospheric states and to sample the sensitivity of the vortex structure to the vertical distribution of parametrized forcing, all simulations using different coefficients were repeated twice with additional vertical restrictions on parametrized tendencies. In one ensemble, all temperature tendencies from parametrized processes were set to zero before calculating advection from the lowest model level up to level 15 (close to 400 meters above sea level), while in the other ensemble, the same tendencies were set to zero from level 15 to the model top.

S2 Pre-processing of CM1 outputs

Given that all simulations use identical initial conditions, we retained only outputs after 80 hours of model integration. All retained hourly model outputs were used directly to construct a machine learning training dataset composed by the fields of PV, potential temperature, and tangential wind speed. Figure S1 shows representative tangential wind speed and PV fields, randomly chosen from 10 groups of the training dataset, formed by k-means clustering applied to all PV fields.

To form the training dataset, all fields were linearly interpolated from native model levels onto isobaric levels spanning 1000 to 100 hPa. The pressure grid was constructed in three segments to provide enhanced resolution in the lower and middle troposphere: 10 hPa spacing from 1000 to 700 hPa, a 20 hPa spacing from 700 to 300 hPa, and a 25 hPa spacing from 300 to 100 hPa. Missing values below the ocean

surface were filled using a biharmonic image inpainting algorithm (inpaint_biharmonic from Python scikit-image library) which solves a bi-Laplacian partial differential equation over the masked region subject to the known boundary conditions of non-NaN values, yielding a smooth, physically reasonable continuation of the field into the sub-surface region (roughly from 1000 to 950 hPa in Fig. S2). Interpolation to pressure levels, rather than using the native model vertical coordinate expressed as distance above sea level, was employed to restrict the U-Net input and output variables to those required for potential vorticity (PV) calculation, namely potential temperature (θ) and tangential wind speed. Because the PV equation is incorporated into the U-Net loss function, all predicted variables must explicitly appear in the PV formulation. Otherwise, a physics-informed U-Net trained on native model levels would also require the prediction of additional quantities, such as air density, to ensure consistency in the PV calculation.

S3 An ensemble of neural network configurations

PV inversion was performed using U-Net convolutional neural networks (Ronneberger et al. 2015). Each network receives a single-channel 2D PV field defined on a pressure-radius grid and predicts the corresponding tangential wind and potential temperature perturbation fields simultaneously, yielding a two-channel output. Two architectural variants were developed and evaluated, referred to hereafter as V1 and V2; these are described in turn below and summarized in Table S1. In total, 101 trained model checkpoints were produced and evaluated, decomposed into 85 models belonging to the V1 configuration and 16 models belonging to V2.

Architectures

V1: U-Net with Dilated Residual Bottleneck: The V1 architecture (UNetDilatedConfigurable) follows the encoder-bottleneck-decoder design of the original U-Net, with skip connections between spatially symmetric encoder and decoder levels. All normalisation layers use Group Normalisation (Wu and He, 2018) with up to 8 groups, and all activations use the Sigmoid Linear Unit (SiLU; Elfwing et al. 2018).

- *Encoder:* The encoder comprises $D+1$ levels ($D = 3$ in the reference configuration). Each level applies two successive convolution blocks, each consisting of a 3×3 convolution, Group Normalisation, and SiLU activation. The number of feature channels doubles at each level, starting from a base width of C channels ($C = 16$ in the reference), yielding channel counts $[C, 2C, 4C, \dots, 2^D \cdot C]$. Downsampling between levels is performed by a learned 2×2 stride-2 convolution rather than pooling. In a variant configuration (labeled EARLYSTRIDE1; Table S1), an additional stride-2 convolution is placed at the entry of the first encoder level, halving the spatial resolution before the main encoder sequence begins; bilinear upsampling is applied at the output to restore the original resolution.

- *Bottleneck:* The bottleneck processes the deepest encoder feature map through a sequence of dilated residual blocks flanked by two single convolution blocks. Each dilated residual block applies two 3×3 dilated convolutions (each followed by Group Normalisation), adds the residual, and applies SiLU. The dilation sequence $\{d_1, d_2, \dots\}$ expands the effective receptive field progressively; the reference configuration uses dilations $\{1, 2, 4, 8\}$, covering spatial scales up to 15 grid points per bottleneck block.

- *Decoder:* The decoder mirrors the encoder over D levels. At each level, the feature map is upsampled by a 2×2 transposed convolution, concatenated with the corresponding encoder skip connection (centre-cropped to match spatial dimensions where necessary), and processed by two successive convolution

blocks identical in structure to those of the encoder. A final 1×1 convolution projects the base-width feature map to the two output channels.

- *Input padding*: To ensure compatibility with the power-of-two downsampling factor, the input field is symmetrically padded (replication padding) to the nearest multiple of $2^{(D+1)}$ (when early stride-2 is active) or 2^D otherwise, and the corresponding border is removed from the output.

V2: U-Net with Depthwise-Separable Convolutions and Squeeze-and-Excitation

The V2 architecture (UNetDilatedV2) retains the overall topology of V1 but introduces two modifications aimed at improving parameter efficiency and channel-wise feature recalibration:

- First, every standard 3×3 convolution in the encoder, decoder, and bottleneck is replaced by a depthwise-separable convolution: a depthwise 3×3 convolution (one filter per input channel) followed by a pointwise 1×1 convolution. This factorisation reduces the number of parameters per convolution layer by a factor approximately equal to the number of input channels relative to the number of 3×3 kernel weights, while preserving the spatial mixing capability.
- Second, a Squeeze-and-Excitation (SE) block (Hu et al. 2018) is inserted at the bottleneck after the dilated block sequence. The SE block globally average-pools the spatial dimensions of the bottleneck feature map, passes the resulting channel descriptor through two fully connected layers with a SiLU intermediate activation and a sigmoid output, and multiplies the result back into the feature map channel-wise. The reduction ratio within the SE block is 4, giving a hidden dimension of $C \cdot 2^D / 4$. The dilation sequence is reduced to $\{2, 8\}$ in V2, as the two depthwise-separable dilated blocks were found sufficient for bottleneck coverage. The V2 configuration also activates the early stride-2 encoder entry by default.

Training procedure

Loss function: All models were trained to minimise a composite loss

$$L = L_{sup} + \lambda_{PV} \cdot L_{PV} + L_{MS}$$

- where L_{sup} is the mean squared error (MSE) between the predicted and reference wind and potential temperature fields as it would be the case in a supervised training procedure; L_{PV} is a physics-consistency term computed as the pressure-level-weighted MSE between the PV field re-diagnosed from the predicted (v, θ) fields via the discrete axisymmetric PV operator and the reference PV field re-diagnosed from the target (v, θ) fields on the same pressure grid. L_{MS} provides auxiliary multi-scale supervision by computing L_{sup} at spatially downsampled (average-pooled) resolutions. In the reference configuration, L_{MS} uses two additional scales (factors of 2 and 4) with weights 0.5 and 0.25, respectively. The physics loss weight $\lambda_{PV} = 1.0$ in all configurations except the λ_{PV} ablation series.

- Two-phase training schedule*: Training was divided into two phases to prevent the physics-consistency term from dominating gradient updates before a reasonable supervised solution is established. During the first N_s epochs ($N_s = 5$ in the reference configuration), λ_{PV} was set to zero and only L_{sup} and L_{MS} were active. From epoch $N_s + 1$ onwards, the full composite loss was applied.

Optimiser and hyperparameters: All models were optimised with AdamW (Loshchilov and Hutter 2019) with a learning rate of 2×10^{-3} and weight decay of 10^{-4} , using a batch size of 6 for 80 epochs. The training set comprised 90% of the available samples, with the remaining 10% reserved for validation; the

120 checkpoint with the lowest validation loss was retained. V1 models used a constant learning rate schedule; V2 used cosine annealing to a minimum rate of 10^{-5} together with gradient clipping at unit norm. The random seed was fixed at 42 for all models; three models (SEED1, SEED2, and LR; see Table S1) were trained under identical configuration to the reference to assess stochastic training variability.

125 *Input normalisation:* PV input fields were standardised using the per-channel mean and standard deviation computed over the training set. Output fields were denormalised at inference using the corresponding stored statistics.

Piecewise Fine-Tuning: To enforce the additive linearity property required for piecewise PV inversion namely, that the sum of the network's responses to two complementary sub-fields recovers the response to the full field—base models were subsequently fine-tuned for additional epochs at a reduced learning rate of 5×10^{-4} . Four decomposition strategies were available:

- (i) Amplitude split 60/40: The full PV field (q) is partitioned into $0.6 \cdot q$ and $0.4 \cdot q$, and the combined loss penalises the deviation of $f(0.6 \cdot q) + f(0.4 \cdot q)$ from $f(q)$. Fine-tuning duration: 40 additional epochs.
- 135 (ii) Amplitude split 80/20: As (i), with fractions $0.8 \cdot q$ and $0.2 \cdot q$, providing a more asymmetric decomposition. Fine-tuning duration: 40 additional epochs.
- (iii) Sign-based decomposition: The PV field is decomposed into its positive part $q^+ = \max(q, 0)$ and negative part $q^- = \min(q, 0)$, enforcing $f(q^+) + f(q^-) = f(q)$. Fine-tuning duration: 40 additional epochs.
- 140 (iv) Vertical half-level decomposition: The PV field is set to zero above (below) the mid-pressure index to form the lower (upper) anomaly, and the additive constraint $f(q_{lower}) + f(q_{upper}) = f(q)$ is imposed with additional loss weights $\lambda_{add,sup} = 0.5$, $\lambda_{add,PV} = 0.25$, $\lambda_{add,MS} = 0.25$. Fine-tuning duration: 80 additional epochs.

V1 fine-tuning models (64 fine-tuned checkpoints from 21 base V1 models): All 21 V1 base configurations received three fine-tuning variants: strategies (i), (ii), and a chained variant in which the checkpoint produced by strategy (i) was further fine-tuned with strategy (iii). This gives 63 fine-tuned V1 runs. In addition, the EARLYSTRIDE1 configuration received a fourth fine-tuned variant using strategy (iv), yielding 1 additional model.

150 V2 fine-tuning runs (15 fine-tuned checkpoints from 1 base V2 model): The single V2_DS_SE base model first received all four fine-tuning strategies independently. Subsequently, 11 ordered-pair combinations of distinct strategies were applied sequentially, i.e., the checkpoint produced by one strategy was used as the starting point for a second strategy.

S4 Evaluation of Piecewise Linearity and Model Selection

A prerequisite for piecewise PV inversion is that the trained models should satisfy an additive property. Specifically, for a decomposition of the PV field into two components, $q = q1 + q2$, the potential temperature and tangential wind speed inferred from each individual PV anomaly ($q1$ and $q2$) should sum to the fields inferred from the total PV field (q). That is, if Θ_q and U_q denote the potential temperature and tangential wind speed inferred from the total PV field, the inference should approximately satisfy

$$\Theta_{q1} + \Theta_{q2} \approx \Theta_q,$$

and

160 $U_{q1} + U_{q2} \approx U_q.$

To assess U-Net performance, Fig. S2 presents scatterplots of the 90th percentile of the absolute bias for the inferred potential temperature and tangential wind speed from each of the 101 trained U-Net models: first, we show the 90th percentile of the biases from all grid points, defined as $|\Theta_q - \theta_{WRF}|$ and $|U_q - u_{WRF}|$, that is, the differences between the fields inferred by the U-Net using the full PV field as calculated by the potential temperature and tangential wind speed provided by WRF minus the same θ_{WRF} and u_{WRF} fields (red dots in Fig. S2). This comparison evaluates whether the U-Net can credibly reproduce the atmospheric state of the tropical cyclone. Second, Fig. S2 shows the 90th percentile of the biases associated with the additive property, defined as $|\Theta_{q1} + \Theta_{q2} - \Theta_q|$ and $|U_{q1} + U_{q2} - U_q|$ (blue dots in Fig. S2). The two PV anomalies correspond to distinct regions of the full PV field. The first is defined by the total PV that overlaps with the negative diabatic PV-tracer values within Usagi's inner core (discussed in section 3.3 as inner-core negative PV_{diab} bulb, while the second comprises the remainder of total PV within the domain.

The majority of the 101 trained U-Net models were able to infer Θ_q and U_q from the total PV field, with the red dots in Fig. S2 indicating average 90th-percentile absolute biases of approximately 10 m s^{-1} and 9 K , respectively. These 90th-percentile values represent the more extreme errors within the cyclone structure rather than domain-wide mean biases. The magnitude of these biases is comparable to those obtained using the analytical method of Davis and Emanuel (1991) (Flaounas et al., 2021). They are also not unexpected, given that the U-Net models were trained to reproduce the relationship between PV, wind, and temperature in CM1 simulations, which represent cyclones within a two-dimensional axisymmetric framework. In contrast, the U-Net models are applied here to axisymmetric averages derived from WRF simulations and are evaluated against the corresponding axisymmetric averages of potential temperature and tangential wind speed, introducing an additional source of discrepancy.

On the other hand, all 22 base (non-piecewise fine-tuned) models yielded 90th-percentile $|\theta_{q1} + \theta_{q2} - \theta_q|$ values exceeding 300 K . This is a consequence of the nonlinear approach of a U-Net: the encoder, bottleneck, and skip connections jointly represent global context derived from the full PV distribution. When a partial sub-field is presented in isolation, the network operates outside its training distribution. The magnitude of these errors confirms that supervised training on full PV fields alone confers no piecewise linearity. Fine-tuning with the 60/40 decomposition (strategy i), as described above, reduced errors by roughly a factor of two, but 90th-percentile of θ bias remained above 150 K for all tested models. It is likely that both sub-fields ($0.6 \cdot q$ and $0.4 \cdot q$) lie within the amplitude range seen during training of the 22 base U-Net models. Therefore, enforcing additivity for two large-amplitude inputs competes directly with the supervised reconstruction objective, and the fine-tuning converges to a compromise that satisfies neither constraint. Fine-tuning based on vertical half-level decomposition (strategy iv) similarly produced errors above 150 K , likely because half-level PV anomalies differ structurally from the full-column patterns on which the network was trained, again placing inputs outside a reliable operating regime.

On the other hand, fine-tuning with the 80/20 decomposition produced smaller bias 90th percentile metrics: five models achieved 90th-percentile additivity errors below 5 K in θ and below 5 m s^{-1} of in tangential wind (outlined in bold letters in Table S1). The improved performance relative to the 60/40 split can be probably attributed to the fact that a minor sub-field ($0.2 \cdot q$) is small enough to act as a

perturbation around the network's original training field (which is closer to $0.8 \cdot q$). These five models were retained for all subsequent piecewise PV inversion in section 3.3.

In this study, our primary interest lies in quantifying the contribution of different cyclone substructures to cyclone intensity. Consequently, we focus on comparing the tangential wind speed inferred from the total PV field, U_q , with the sum of the tangential wind speeds inferred from the two PV anomalies, $U_{q1} + U_{q2}$. Figure S3 provides further insight into the ability of the five retained U-Net models to reproduce the tropical cyclone wind field by showing their mean inferred tangential wind speed associated with the two PV anomalies discussed in section 3.3, along with the mean bias, defined as $U_q - (U_{q1} + U_{q2})$. The results reveal a localized region of negative bias for Fig. S3a, with a minimum of approximately -15 m s^{-1} near the cyclone center and within the lower pressure levels, while even smaller bias, of less than 10 m s^{-1} , is found in the case of the second PV anomaly in Fig. S3b. This indicates a spatially confined departure from perfect additivity with negligible biases close to the area of maximum wind speed.

Model label	Arch.	C	D	Dilations	Early stride	λ_{PV}	N_s	Multi- scale scales	Multi- scale weights
light_pvcalc_input	V1	16	3	{1,2,4,8}	off	1.0	5	{2,4}	{0.50, 0.25}
LAMBDA_PV0.1	V1	16	3	{1,2,4,8}	off	0.1	5	{2,4}	{0.50, 0.25}
LAMBDA_PV0.3	V1	16	3	{1,2,4,8}	off	0.3	5	{2,4}	{0.50, 0.25}
LAMBDA_PV0.6	V1	16	3	{1,2,4,8}	off	0.6	5	{2,4}	{0.50, 0.25}
SUP_ONLY_EPO CHS0	V1	16	3	{1,2,4,8}	off	1.0	0	{2,4}	{0.50, 0.25}
SUP_ONLY_EPO CHS3	V1	16	3	{1,2,4,8}	off	1.0	3	{2,4}	{0.50, 0.25}
SUP_ONLY_EPO CHS10	V1	16	3	{1,2,4,8}	off	1.0	10	{2,4}	{0.50, 0.25}
MULTI1	V1	16	3	{1,2,4,8}	off	1.0	5	disabled	—
MULTI2	V1	16	3	{1,2,4,8}	off	1.0	5	{2,4,8}	{0.70, 0.50, 0.30}
MULTI3	V1	16	3	{1,2,4,8}	off	1.0	5	{2,4,8}	{0.50, 0.30, 0.20}
EARLYSTRIDE1	V1	16	3	{1,2,4,8}	on	1.0	5	{2,4}	{0.50, 0.25}
BASECH1	V1	24	3	{1,2,4,8}	off	1.0	5	{2,4}	{0.50,

									0.25}
BASECH2	V1	16	4	{1,2,4,8}	off	1.0	5	{2,4}	{0.50, 0.25}
BOTTLE1	V1	16	3	{1,2,4}	off	1.0	5	{2,4}	{0.50, 0.25}
BOTTLE2	V1	16	3	{1,2,4,8,16}	off	1.0	5	{2,4}	{0.50, 0.25}
LR	V1	16	3	{1,2,4,8}	off	1.0	5	{2,4}	{0.50, 0.25}
SEED1	V1	16	3	{1,2,4,8}	off	1.0	5	{2,4}	{0.50, 0.25}
SEED2	V1	16	3	{1,2,4,8}	off	1.0	5	{2,4}	{0.50, 0.25}
balanced_bestbet	V1	16	3	{1,2,4,8}	off	0.6	5	{2,4,8}	{0.70, 0.50, 0.30}
fieldfirst_robust	V1	16	3	{1,2,4,8}	off	0.3	10	{2,4,8}	{0.50, 0.30, 0.20}
physics_aggressive	V1	16	3	{1,2,4,8}	on	1.0	3	{2,4,8}	{0.70, 0.50, 0.30}
V2_DS_SE	V2	16	3	{2,8}	on	1.0	5	{2,4}	{0.50, 0.25}

215 **Table S1.** Summary of base model configurations. All V1 runs share: architecture V1, $C = 16$, $D = 3$,
 early stride-2 = off, $\lambda_{PV} = 1.0$, $N_s = 5$, multi-scale scales {2,4} with weights {0.5, 0.25}, AdamW with LR
 = 2×10^{-3} and weight decay = 10^{-4} , 80 epochs, batch size 6, seed 42, unless stated otherwise. (*)
 independent replicate with identical hyperparameters; (**) V2 architecture. Lines in bold depict the U-
 220 Net models fine-tuned with 80/20 strategy iv (see text) which have been used for the analysis in section
 3.3 of the main article.

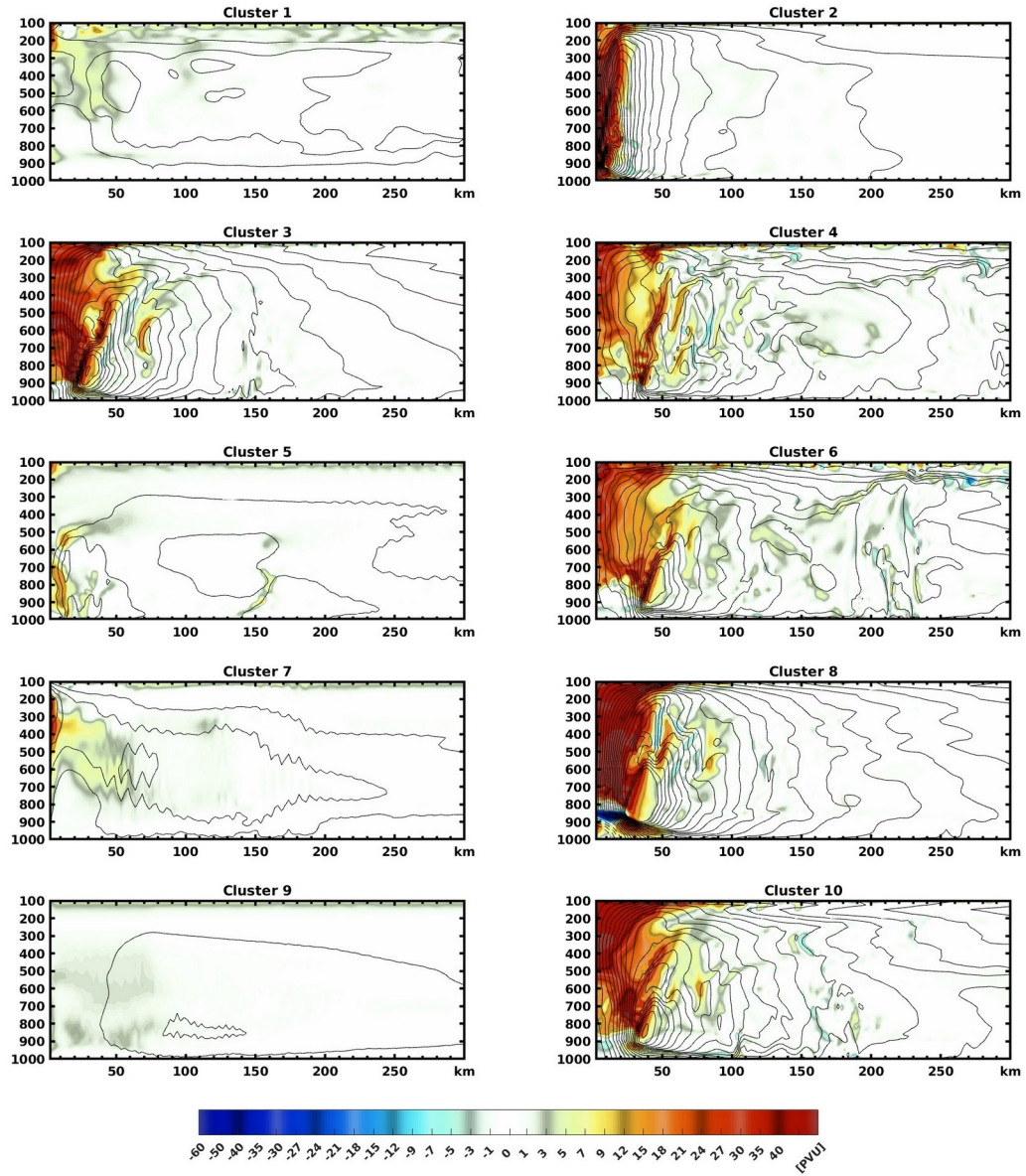


Figure S1 Fields of PV (in colour) and tangential wind speed (contours with an interval of 5 m s^{-1}) from the CM1 outputs that compose the training dataset. Fields correspond to random members of each of the 10 clusters created applying the k-means clustering approach to all PV fields.

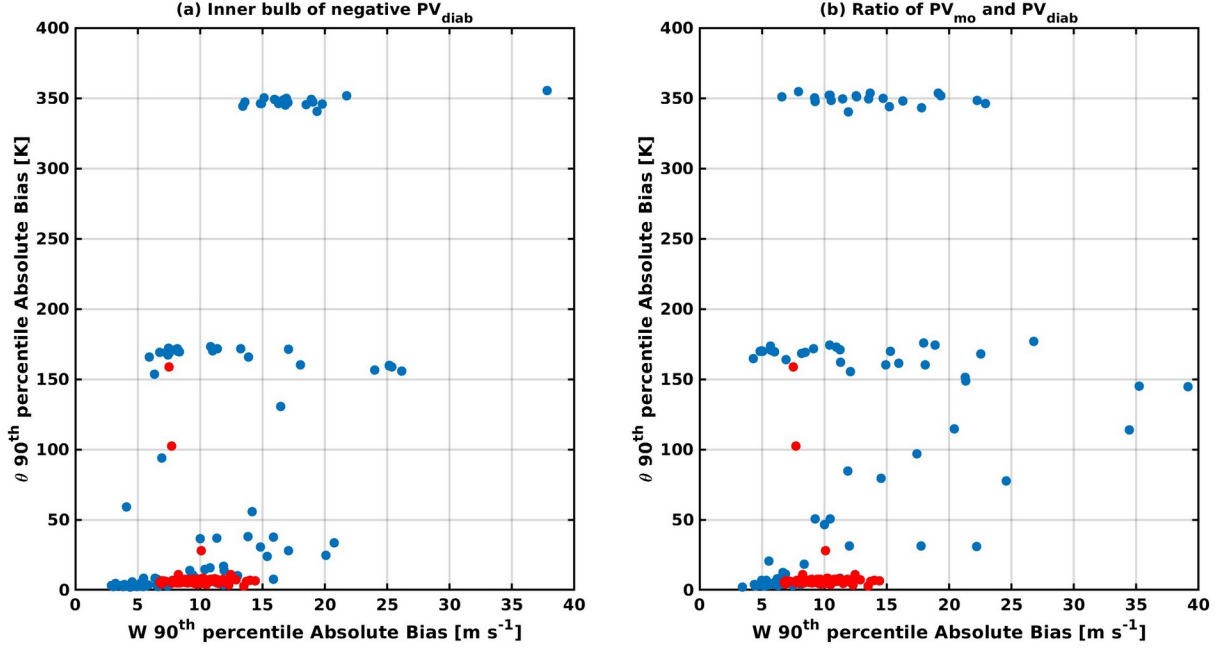


Figure S2 Each dot represents the 90th percentile of the absolute bias for one of the 101 trained U-nets within the inference domain, shown for potential temperature (y-axis) and tangential wind speed (x-axis) for two different PV inversions (see Section 3.3): **(a)** for the inner bulb of negative PV_{diab} and **(b)** for the approximated PV_{mo} and PV_{diab} components. Red dots denote the bias between the radially averaged WRF fields and the fields inferred from the total PV derived from those WRF fields. Blue dots denote the bias between the sum of the fields inferred from the individual PV anomalies (see text) and the fields inferred from the total PV.

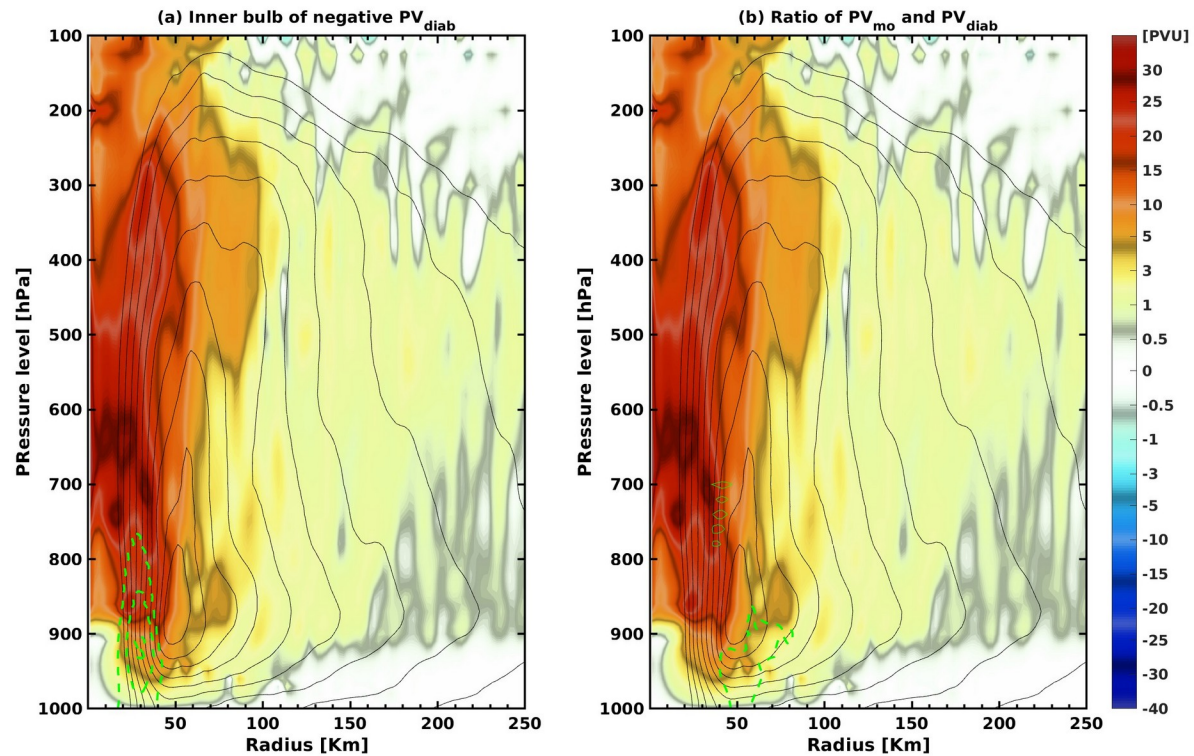


Figure S3 Vertical-radial distribution of total PV in TC Usagi at mature stage (in colour). Black contours (interval of 5 m s^{-1} starting from 20 m s^{-1}) show the five U-Net ensemble-mean inferred tangential wind speed associated with the total PV (in colour). Solid (dashed) green contours show positive (negative) mean wind bias with an interval of 5 m s^{-1} , with respect to the inferred wind speed from total PV (in colour) for two different PV inversions (see Section 3.3): **(a)** for the inner bulb of negative PV_{diab} and **(b)** for the approximated PV_{mo} and PV_{diab} components.

References

Bryan, G. H.: Effects of Surface Exchange Coefficients and Turbulence Length Scales on the Intensity and Structure of Numerically Simulated Hurricanes, *Mon. Wea. Rev.*, 140, 1125–1143, <https://doi.org/10.1175/MWR-D-11-00231.1>, 2012.

Elfving, S., Uchibe, E., and Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning, *Neural Networks*, 107, 3–11, <https://doi.org/10.1016/j.neunet.2017.12.012>, 2018.

Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, <https://doi.org/10.48550/ARXIV.1711.05101>, 2017.

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E.: Squeeze-and-Excitation Networks, <https://doi.org/10.48550/ARXIV.1709.01507>, 2017.

255 Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, vol. 9351, edited by: Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., Springer International Publishing, Cham, 234–241, https://doi.org/10.1007/978-3-319-24574-4_28, 2015.

Wu, Y. and He, K.: Group Normalization, <https://doi.org/10.48550/ARXIV.1803.08494>, 2018.