Supplementary Information

Supplementary Information

Sensitivity analyses

These five experiments support the main-text claim that, once the task is decomposed correctly, performance is more strongly influenced by the available data and physical forcings than by parameter count alone. Unless stated otherwise, each experiment uses the best model for the relevant stage with the hyperparameters from the tuning sweep.

Hyperparameter tuning: deterministic architectures benefit monotonically from capacity; probabilistic models require balancing capacity against training stability

The tuning protocol is described in Methods; here we report what the sweep revealed about how each architecture responds to capacity. For the seven deterministic super-resolution architectures (MLP, CNN, EDSR, RCAN, HREDSR, U-Net, SwinIR), validation nRMSE decreases close to monotonically with capacity across the tested range, so the largest feasible configuration is generally selected; U-Net is a partial exception, showing training instability (abrupt loss and nRMSE spikes) at some intermediate configurations rather than a smooth capacity–performance curve. For the probabilistic generative models and the adversarially trained GAN, validation loss is a noisier proxy for the metric of interest: for EDM, the dHR→HR configuration was chosen for consistent convergence across seeds rather than the single lowest validation loss, since a smaller model scored marginally lower but less reliably; SwinIR's embedding dimension was likewise fixed across all three prediction tasks at its dHR→HR-optimal value rather than re-tuned per stage (Supplementary Table S6).

Table 1: Supplementary Table S6 | Hyperparameter tuning sweep. Swept hyperparameter ranges and the selected configuration per architecture and prediction stage, with validation-set metrics at that configuration (best of up to three seeds by validation loss). "Width"/"channels" denotes the hidden feature width (per-layer channel count); "blocks" denotes the number of residual/convolutional blocks; "groups" denotes the number of residual groups (RCAN only); for SwinIR, "blocks/group" denotes the number of Swin Transformer blocks within each of its four fixed transformer-block groups. nRMSE: normalized root mean square error (%); SSIM: structural similarity index.

Architecture	**Swept hyperparameters**	**Stage**	**Selected configuration**	**Val. nRMSE (%)**	**Val. SSIM**
MLP	2-layer MLP, width $\in \{4, 8, 16, 32\}$	dHR→HR	$(32, 32)$, residual	3.98	0.9744
		LR→dHR	$(32, 32)$, residual	17.71	0.8381
CNN	width $\in \{32, 64, 128, 256\}$	dHR→HR	256, residual	3.30	0.9815
		LR→dHR	256, no residual	15.31	0.8742
EDSR	width $\in \{32, 64, 128\}$, blocks $\in \{2, 4, 8, 16\}$	dHR→HR	128×16, residual	2.47	0.9889
		LR→dHR	128×16, no residual	14.06	0.8946
RCAN	width $\in \{32, 64, 128\}$, blocks/groups $\in \{2, 4\}$	dHR→HR	128×4×4, residual	2.42	0.9891
		LR→dHR	128×4×4, residual	14.04	0.8972
HREDSR	width $\in \{32, 64, 128\}$, blocks $\in \{2, 4, 8, 16\}$	dHR→HR	128×16, residual	2.31	0.9904
		LR→dHR	128×16, residual	13.88	0.8973
U-Net	channels $\in \{32, 64, 128\}$, blocks $\in \{2, 4\}$	dHR→HR	128×4, residual[a]	2.10	0.9910
		LR→dHR	128×4, residual	13.81	0.8730
SwinIR[d]	embed. dim $\in \{48, 96, 128\}$, blocks/group $\in \{2, 4, 6\}$ (4 groups)	dHR→HR	128, 6 blocks/group, residual	2.36	0.9897
		LR→dHR	128[b], 6 blocks/group, residual	14.04	0.8963
GAN	adv. weight $\in \{.05, .1, .15, .2\}$, disc. lr $\in \{3, 5, 8\} \times 10^{-4}$	dHR→HR	adv=0.05, lr=8e-4, residual	2.31	0.9905
		LR→dHR	adv=0.05, lr=5e-4, no residual	14.14	0.8930
Flow-matching	channels $\in \{32, 64, 128\}$, blocks $\in \{2, 4, 6\}$	dHR→HR	128×4, residual	2.71	0.9862
		LR→dHR	128×4, residual	15.85	0.8488
EDM	channels $\in \{32, 64, 128\}$, blocks $\in \{2, 4, 6\}$	dHR→HR	64×4, residual[c]	2.88	0.9844
		LR→dHR	64×4, residual	16.09	0.8574

[a]U-Net shows training instability at some intermediate configurations, so its capacity ordering is not fully monotonic. [b]The embedding dimension was fixed across all three prediction tasks (LR→dHR, dHR→HR, LR→HR) at its dHR→HR-optimal value rather than re-optimized per stage. [c]Chosen for convergence robustness across seeds rather than minimum validation loss (32 channels, 6 blocks scored marginally lower: 0.64 vs. 0.66 validation loss). [d]SwinIR's attention-head count (4 per block group) and local attention window size (8) were held fixed throughout the sweep, not tuned.

Coarse-cell consistency constraints act as exact guarantees and regularizers

We trained every architecture on the super-resolution stage (dHR→HR) with each of four constraint settings: none, additive, multiplicative, and softmax [1]. Beyond guaranteeing coarse-cell concentration consistency, the constraint also regularizes training, with the largest improvement for the smallest models (the MLP and plain CNN), whose unconstrained outputs are otherwise the least well-behaved. The multiplicative form is the best or tied-best default for most architectures, but the optimal choice is architecture-dependent (softmax is best for EDM, additive for the U-Net) and softmax destabilizes several models, so the constraint should be matched to the architecture rather than applied blindly (Supplementary Fig. S1).

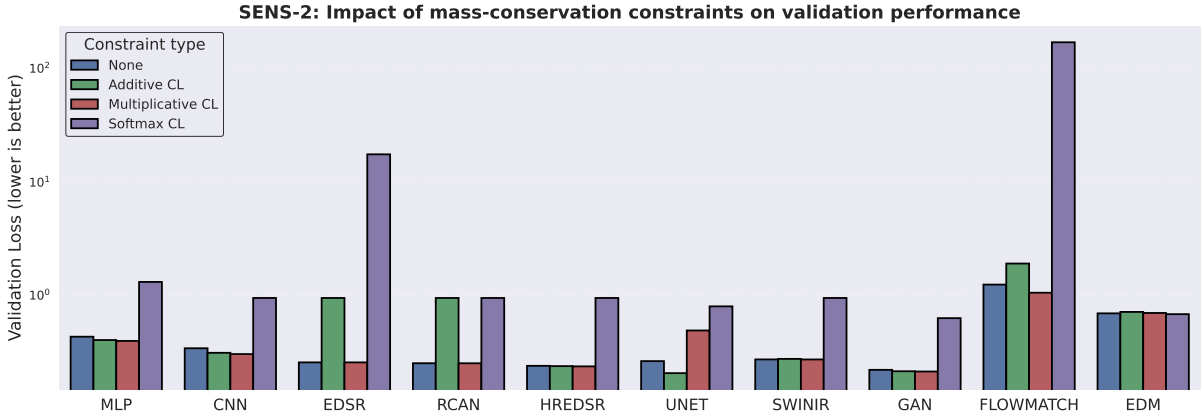

Figure 1: Supplementary Figure S1 | Coarse-cell consistency constraints across architectures. Validation loss for each architecture on the dHR→HR stage under four constraint settings (none, additive, multiplicative, softmax).

Meteorology is the dominant ancillary input

We measured the contribution of each ancillary group (meteorology, emissions, orography, land use) by adding and removing them one at a time, for both stages (EDM on LR→dHR and SwinIR on dHR→HR), over five seeds. Meteorology captures the vast majority of the achievable ancillary benefit; emissions and orography contribute modestly and comparably, and land use contributes negligibly (Supplementary Fig. S2). The limited role of emissions is likely specific to our 2× scale factor: at larger scale-ups, where more sub-grid emission heterogeneity is unresolved in the coarse field, we would expect their contribution to grow.

Univariate models are only marginally better than a joint model

We compared a single multivariate SwinIR (all four pollutants) against four univariate SwinIR models (one per pollutant) on the super-resolution stage. Training a separate model per pollutant lowers nRMSE by only 2–10% (largest for PM_{10}), which does not justify the added training and deployment cost of one model per species (Supplementary Table S3).

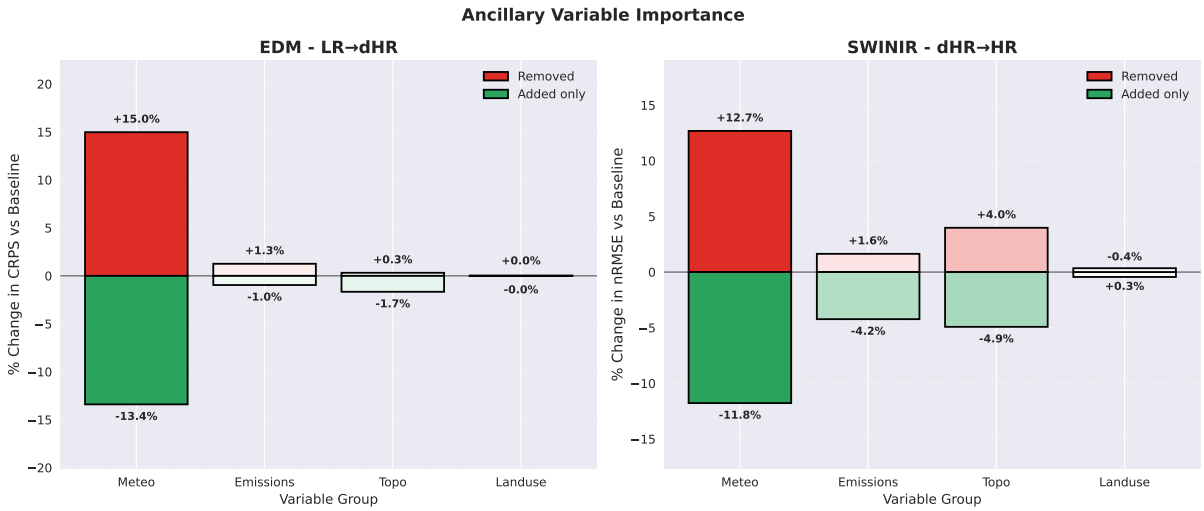

Figure 2: Supplementary Figure S2 | Ancillary-variable importance. Change in performance when each ancillary group is added or removed, for both stages.

Table 2: Supplementary Table S3 | Univariate vs. multivariate model performance (Stage 2, SwinIR). Test set metrics and percentage changes relative to multivariate baseline. Purple indicates improvement; orange indicates degradation (accounting for metric polarity). nRMSE: normalized root mean square error (%); R: Pearson correlation; pMAE: percentage mean absolute error; SSIM: structural similarity index.

Pollutant	Config	nRMSE (%)	R	pMAE	SSIM
NO_2	Multivariate	5.64	0.9993	0.0531	0.9948
	Univariate	5.36	0.9994	0.0505	0.9956
	Δ (%)	–4.9%	+0.01%	–5.0%	+0.08%
O_3	Multivariate	0.83	0.9995	0.0244	0.9770
	Univariate	0.81	0.9995	0.0251	0.9780
	Δ (%)	–2.2%	+0.002%	+3.0%	+0.10%
PM_{10}	Multivariate	3.74	0.9993	0.1439	0.9871
	Univariate	3.36	0.9994	0.1359	0.9880
	Δ (%)	–10.2%	+0.008%	–5.6%	+0.09%
$PM_{2.5}$	Multivariate	2.15	0.9993	0.0258	0.9791
	Univariate	1.99	0.9994	0.0231	0.9817
	Δ (%)	–7.5%	+0.009%	–10.6%	+0.27%

Spatial generalization degrades under geographic holdout

We split the Iberian domain into quadrants, trained SwinIR on three, and tested on the held-out quadrant, repeating for each quadrant. To assess performance degradation, validation performance of these models is compared to a reference model trained on a 75% random sample of a full domain training set (including samples from all four quadrants). Results show that performance degrades by 12–66% across holdouts (averaging $\sim 40\%$), and most for $PM_{2.5}$, indicating that the models depend partly on region-specific spatial relationships [2]. Transfer to other regions with contrasting terrain or climate will therefore need dedicated training data or domain adaptation (Supplementary Fig. S4).

Performance primarily follows total sample number within the tested ranges

We scaled the training data along three axes independently: the fraction of the dataset, the number of years, and the number of ensemble members. Adding years and adding members lie on the same curve: only the total number of training samples matters, not how that total is reached (Supplementary Fig. S5).

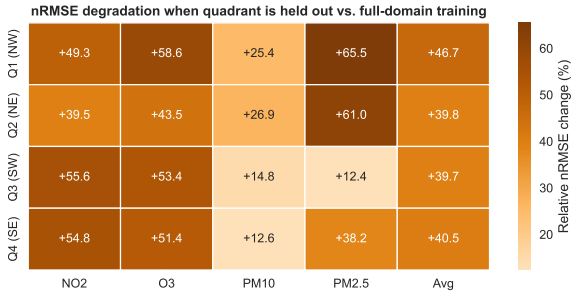

Figure 3: Supplementary Figure S4 | Spatial generalization under quadrant holdout. Performance degradation when testing on a held-out geographic quadrant, per pollutant.

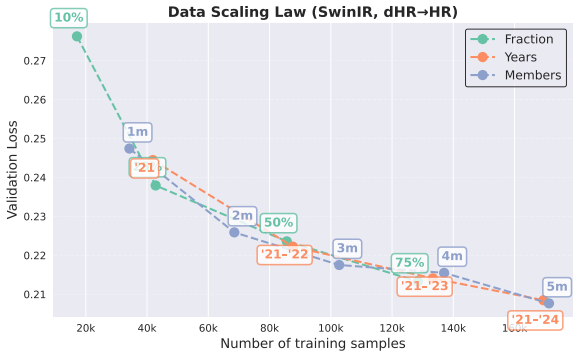

Figure 4: Supplementary Figure S5 | Data-scaling behavior. Performance as a function of training-set size along three axes (samples, years, ensemble members).

Reference tables

The tables below record the configuration and parameter settings referred to in Methods.

Table 3: Supplementary Table S7 | MONARCH run configuration. Configuration of the two independent MONARCH integrations underlying the paired dataset.

	LR (10 km)	**HR (5 km)**
Horizontal grid spacing	0.1° rotated (≈10 km)	0.05° rotated (≈5 km)
Grid dimensions, raw	121 × 141	241 × 281
Grid dimensions, used (cropped)	110 × 130	220 × 260
Domain extent, used (cropped)	34.0–45.2°N, −12.8–5.4°E	34.0–45.3°N, −12.8–5.5°E
Vertical layers	24	
Model top	50 hPa	
Lateral boundary conditions	ERA5 (meteorology), CAMS global forecast on pressure levels (composition); identical for both runs	
Model version	MONARCH v2.9.2	
Output frequency	Hourly	

Table 4: Supplementary Table S8 | Training and sampling hyperparameters. Loss function, optimizer settings, batch size, epoch budget, and early-stopping rule used for the main architecture-comparison experiments (Stage 1, Stage 2, and the one-stage baselines). All families use AdamW [3] ($\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 10^{-2}) with a cosine-annealing schedule decaying to 1% of the initial learning rate, and early stopping on the validation loss with a patience of 5 epochs without improvement. λ_{l1}, λ_{grad}: L1Grad loss weights.

Architecture family	Loss	Learning rate	Batch size	Epochs
MLP, CNN, EDSR, RCAN, HREDSR	L1Grad ($\lambda_{l1} = 1.0$, $\lambda_{grad} = 0.2$)	5×10^{-4}	16	30
GAN	L1 ($\lambda = 1.0$) + adversarial[a]	5×10^{-4}	16	30
SwinIR, U-Net[b]	L1Grad ($\lambda_{l1} = 1.0$, $\lambda_{grad} = 0.2$)	5×10^{-4}	4	30
Flow-matching, EDM	MSE (velocity / denoising score)	5×10^{-4}	8	100

[a]Adversarial weight and the WGAN-GP discriminator's own learning rate are swept and reported per stage in Supplementary Table S6.
[b]Batch size reduced relative to the other architectures because of memory footprint, not a tuned choice.

Table 5: Supplementary Table S9 | WHO 2021 guideline values used for threshold exceedance. Nominal averaging period is the guideline's own definition; the evaluation itself compares these values against native hourly concentrations (see Methods).

Pollutant	WHO 2021 guideline value	Nominal averaging period
NO_2	10 $\mu g\,m^{-3}$	Annual mean
O_3	60 $\mu g\,m^{-3}$	Peak season (6-month running mean of daily max 8-h mean)
PM_{10}	15 $\mu g\,m^{-3}$	Annual mean
$PM_{2.5}$	5 $\mu g\,m^{-3}$	Annual mean

Supplementary References

[1] Harder, P. *et al.* Generating physically-consistent high-resolution climate data with hard-constrained neural networks. In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning* (2022).

[2] Harder, P. *et al.* Benchmarking the geographic generalization of deep learning models for precipitation downscaling. *Scientific reports* **16**, 3733 (2026).

[3] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)* (2019).