

Does enforcing physics improve evapotranspiration downscaling? A controlled comparison of mass-conserving and direct deep-learning models for WaPOR

Zolo Kiala¹

Z.Kiala@cgiar.org

International Water Management Institute

Karim Bergaoui

ACQUATEC Solutions

Petra Schmitter

International Water Management Institute

Mariangel Garcia Andarcia

International Water Management Institute

Article

Keywords: evapotranspiration downscaling, physics-informed deep learning, super-resolution, zero-shot transfer

Posted Date: September 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10805791/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Does enforcing physics improve evapotranspiration downscaling? A controlled comparison of mass-conserving and direct deep-learning models for WaPOR

Authors: Zolo Kiala^{1*}, Karim Bergaoui², Petra Schmitter³, Mariangel Garcia Andarcia³,

¹International Water Management Institute, Block G Ground Floor, Hatfield Gardens, 333 Grosvenor St, Hatfield, Pretoria, 0083, South Africa ²ACQUATEC Solutions, United Arab Emirates ³International Water Management Institute, 127 Sunil Mawatha, Battaramulla, Sri Lanka 10120

*Correspondence: Z.Kiala@cgiar.org

Abstract

Satellite evapotranspiration products underpin agricultural water accounting. Yet the field-scale (20 m) data irrigation needs exist for only a handful of sites, whereas coarse (≈ 300 m) products span the entire African continent. This motivates deep-learning downscaling of the continental product to field scale. Physics-informed networks that enforce hard conservation constraints are widely advocated, on the premise that they improve accuracy and transfer. Yet this premise is rarely tested in a controlled way. To our knowledge, this is the first controlled test of it for evapotranspiration downscaling. It is also the first to adapt image-restoration and geospatial foundation models to the task, and to transfer across contrasting African agro-climates. We downscale WaPOR actual evapotranspiration (ETa) from Level-1 (≈ 300 m) to Level-3 (20 m). Inputs, training, sites, and protocol are held fixed across ten African irrigation pilots. We compare three direct super-resolution models (a compact SwinIR transformer, a Prithvi Vision Transformer, and their ensemble) against a physics-compliant network (PRSN). Direct models explain 78–91 % of L3 ETa variance and transfer zero-shot to unseen agro-climates. They roughly halve naive-upsampling error and beat an operational per-pixel baseline. A compact 1.8 M-parameter SwinIR matches the 300 M-parameter Prithvi foundation model. The physics-compliant models keep their guarantees but are less accurate at every site, by ≈ 6 % to ≈ 47 % in relative RMSE. Neither a Budyko nor a Penman–Monteith anchor closes the gap. Hard constraints therefore buy consistency, not accuracy. Direct super-resolution is the practical route to field-scale ETa across the whole of Africa.

Keywords: evapotranspiration downscaling; physics-informed deep learning; super-resolution; zero-shot transfer

Highlights

- A controlled, identical-input benchmark of physics-compliant vs direct ET downscaling
- Enforcing mass conservation buys consistency, not accuracy, at every African site

- Direct super-resolution transfers zero-shot to held-out African agro-climates
 - A compact 1.8 M-parameter SwinIR beats the 300 M-parameter Prithvi foundation model
 - Deep models roughly halve naive L1-upsampling error and beat operational CatBoost
-

1. Introduction

Irrigated agriculture is the largest consumptive user of freshwater, accounting for roughly 70 % of global withdrawals (FAO 2017; WWAP 2021), so quantifying agricultural water consumption is central to basin-scale water accounting, drought monitoring, and irrigation-performance assessment (Molden et al. 2010; Karimi et al. 2013). The relevant flux is actual evapotranspiration (ETa), and satellite remote sensing has become the primary means of mapping it at scale (Bastiaanssen et al. 1998; Senay et al. 2013; Anderson et al. 2012). Across sub-Saharan and North Africa, however, irrigation is fragmented and heterogeneous, so water-accounting applications need ETa resolved at the field scale (tens of metres) rather than at the hundreds-of-metres resolution of most operational products (You et al. 2011; Xie et al. 2014).

The FAO WaPOR database is a major advance for data-scarce regions, delivering harmonised dekadal ETa at three nested levels (FAO 2018; Blatchford et al. 2020). Its two extremes define the problem we target. Level-1 (L1, ≈ 300 m) covers the entire African continent every dekad but is too coarse to resolve smallholder fields, whereas Level-3 (L3, 20 m) reaches the agronomically useful resolution but exists only over a handful of managed irrigation schemes. Bridging this gap means producing L3-quality 20 m ETa where only the continental L1 product exists, that is, training where L3 is available and deploying, by transfer, where it is not.

This is fundamentally a data-generation problem, and it is where deep learning becomes indispensable. Field-scale ETa cannot be observed everywhere: high-resolution reference data are expensive to produce and exist for only a handful of managed schemes, while dense in-situ networks (eddy-covariance towers, lysimeters) are virtually absent across most of Africa. Conventional approaches stall in this regime, since they require either ground truth that does not exist or physically-based models whose inputs are themselves unavailable at the field scale.

A model could learn the ETa–covariate relationship where reference data do exist, then predict physically-plausible 20 m ETa where they do not. In principle, this would extend those scarce observations from cheap, globally-available inputs: optical imagery, terrain, season, and coarse ETa. Used this way, AI would act not merely as an interpolator, but as a means of propagating costly Earth-observation reference data into under-monitored regions. We test the transferability such a use demands, training where L3 exists and evaluating zero-shot elsewhere. But note at the outset that all our evaluation sites still have L3, which serves as the yardstick. The genuine no-reference deployment regime is therefore motivated here, not yet validated.

Bridging coarse-to-fine resolution is a super-resolution problem, and two modelling philosophies compete for it. Direct super-resolution treats downscaling as image-to-image regression: a network maps coarse ETa plus high-resolution covariates (optical indices, terrain, season) straight to the 20 m target. Tree-based and convolutional/transformer models in this family capture the non-linear ETa–covariate relationship well

(Singh et al. 2014; Ke et al. 2016; Che et al. 2022). They provide no guarantee, however, that their outputs are physically admissible.

Physics-informed (or physics-compliant) downscaling instead embeds hydrological and energy-balance constraints directly into the network, so that predictions are admissible by construction. The 20 m field must re-aggregate to the coarse observation (mass conservation), remain non-negative, and respect an energy ceiling. Such theory-guided and constraint-enforcing learning is widely advocated. The premise is that hard constraints not only guarantee consistency but also improve accuracy and out-of-domain generalisation. The reasoning is that the model no longer expends capacity on impossible solutions. It also stays anchored when it leaves the training domain (Karpatne et al. 2017; Reichstein et al. 2019; Beucler et al. 2021; Willard et al. 2022).

That premise, however, is rarely tested in a controlled way for ETa downscaling. Physics-informed architectures are typically compared against weaker baselines or on a single site. This confounds the effect of the constraints with differences in inputs, capacity, or training, and transferability across agro-climates is seldom evaluated. This paper subjects the premise to a like-for-like test for WaPOR ETa. Holding the inputs, training data, sites, and evaluation protocol fixed, we ask a single falsifiable question: *Given identical inputs and training data, does enforcing physical constraints improve the accuracy of $L1 \rightarrow 20$ m ETa downscaling, in-domain and under zero-shot transfer?*

To answer it, we benchmark three direct super-resolution models against a physics-compliant network (PRSN). The direct models are a compact SwinIR transformer, a Prithvi-V1 geospatial ViT, and their ensemble. PRSN enforces exact mass closure, non-negativity, an energy ceiling, and an advection gate. It is evaluated in two forms. The first is a Budyko-ratio regime-attention variant designed to aid transfer. The second replaces the energy term with a full FAO-56 Penman–Monteith potential-ET anchor, the strongest and most physically faithful implementation. All models share the same nine 20 m input channels and the same seven-site training set. They are evaluated at six of the ten pilots. Three are in-domain hold-outs on unseen years: Baixo, Lamego, and Tunisia (JEN). Three are fully held-out zero-shot transfer sites: MIT, KOGA, and KMW. Significance is tested per dekad. The remaining four training sites are single-year and yield no hold-out.

To our knowledge, this is also the first study to adapt SwinIR and the Prithvi-V1 geospatial foundation model to satellite ETa downscaling. Both have been applied to optical image restoration and Earth-observation tasks, but neither had previously been repurposed for coarse-to-fine ETa super-resolution. It is also, to our knowledge, the first ETa-downscaling model demonstrated to transfer across contrasting African agro-climates. Rather than being calibrated to a single scheme, it is trained on seven irrigation pilots. It is then shown to generalise zero-shot to three fully held-out sites, spanning Mediterranean North Africa, the East African highlands, and mid-altitude East Africa. This directly addresses the operational reality that L3 reference data exist for only a handful of schemes. Our contribution is therefore threefold. Methodologically, we introduce two strong, previously untested super-resolution backbones to the ETa-downscaling problem. Operationally, we deliver the first downscaler shown to work across African agro-climates by transfer. Conceptually, we provide a controlled answer to whether, and where, physical constraints earn their place, with direct implications for how operational ETa downscaling should be built.

2. Materials and Methods

2.1 The physics-compliant model under test

The physics-compliant architecture (here named **PRSN**, the Physics-Reconciled Scaling Network) implements the leading “Scale-Discrepancy Reconciliation” hypothesis. It decomposes the prediction as

$$\text{ET}_{20} = W \cdot R, \quad R = \text{softplus}(\bar{L}_1 + \delta), \quad (1)$$

where $\bar{L}_1$ is the coarse L1 anchor averaged over each ~300 m block, R is the resulting *reconciled coarse total*, and:

- δ is a per-block *scale-discrepancy correction* predicted from the block mean and variance of the high-resolution covariates. It addresses the systematic L1–L3 magnitude gap that arises from Jensen’s inequality when a non-linear ET operator is applied at coarse resolution (sub-pixel heterogeneity $\Rightarrow$ aggregation bias).
- W is a non-negative 20 m *texture map* whose mean is exactly one inside every coarse block. The block-softmax construction guarantees (i) **exact self-closure** — $\text{block-mean}(\text{ET}_{20}) = R$ exactly, i.e. the 20 m field re-aggregates to the model’s *own reconciled coarse total* R (which, because of the softplus and the learned correction δ , is **not in general equal to the observed L1**) — and (ii) **ET ≥ 0** , fixing the negative-flux flaw of mean-centred formulations. The distinction between closure to R and closure to observed L1 is quantified in §3.3.
- An optional **energy ceiling** (a soft hinge on net-radiation-bounded ET) and a learnable **advection gate** add physical realism while preserving the hard closure.
- An optional **Budyko-ratio regime attention** module conditions δ and the sharpness of W on a multi-scale aridity index (supply from antecedent rainfall, demand from an energy proxy). This is the proposed *transferability* mechanism: it should let the model adapt its bias correction to unseen agro-climates.

The decomposition of Eq. (1) is mathematically attractive and delivers genuine guarantees (Figure 1). The open question, and the subject of this study, is whether these guarantees come for free, or at an accuracy cost.

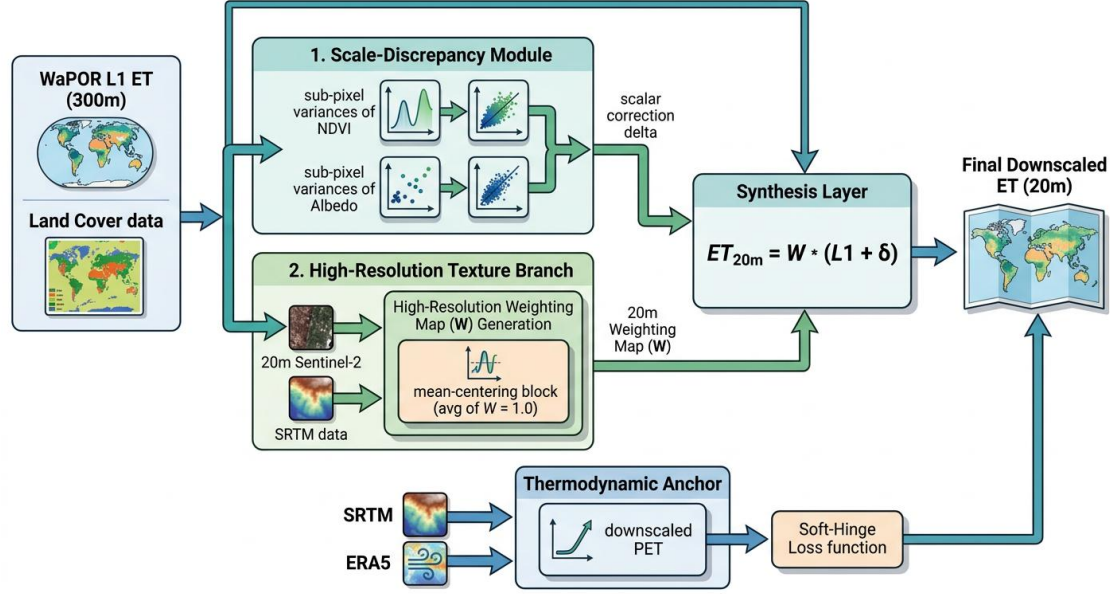


Figure 1. Architecture of the decoupled physics-compliant PRSN network. A scale-discrepancy module predicts the per-block bias correction δ ; a texture branch (Sentinel-2 + SRTM) predicts the 20 m weighting map W (block mean = 1). The synthesis layer forms $ET_{20} = W \cdot R$, with $R = \text{softplus}(\bar{L}_1 + \delta)$, giving exact self-closure to the reconciled coarse total R and non-negativity by construction. A downscaled Penman–Monteith potential evapotranspiration (PET; SRTM + ERA5) supplies a soft-hinge energy-ceiling loss.

2.2 Study areas

Our study uses ten African irrigation pilots spanning the major agro-climatic zones of the continent (Figure 2). They were chosen to stress-test transferability over the widest practical range of climate, cropping system, field size, and irrigation geometry represented in the WaPOR Level-3 archive. Seven contribute to training and three (MIT, KOGA, KMW) are fully held out for zero-shot transfer. Reportable accuracy comes from the six pilots that have a hold-out: the three transfer sites, plus Baixo, Lamego, and Tunisia (JEN) on unseen years. The four single-year training sites (AWA, GEZ, ODN, LOU) add agro-climatic diversity to training but provide no in-domain hold-out.

The site set was selected to expose the models to both local fine-scale structure and continental-scale domain shift. It spans large gravity-irrigated schemes, smaller highland and mid-altitude systems, Mediterranean orchards and field crops, Sahelian rice systems, and sub-humid Southern African schemes. Their characteristics are summarised in Table 1.

Table 1. Characteristics of the ten African irrigation pilot sites (seven training, three zero-shot transfer).

Site	Country	Agro-climate	Cropping system / field geometry	Role	
BAIXO	Mozambique	Semi-arid subtropical (Limpopo, ~25°S)	Sugarcane + smallholder, medium fields	Training (in-domain)	(in-
LAMEGO	Mozambique	Sub-humid (Sofala)	Smallholder, fragmented fields	Training (in-domain)	(in-
AWA	Ethiopia	Semi-arid (Rift Valley)	Sugar estate + horticulture	Training (diversity)	
GEZ	Sudan	Hot arid (Gezira)	Cotton/cereal, very large geometric blocks	Training (diversity)	
ODN	Mali	Sahelian semi-arid (Office du Niger)	Rice-dominated, canal grid	Training (diversity)	
LOU	Morocco	Mediterranean (~35°N)	Irrigated cereal/orchard	Training (diversity)	
JEN	Tunisia	Mediterranean (Medjerda, ~36°N)	Irrigated cereal, large valley fields	Training (diversity)	
MIT	Algeria	Mediterranean (~36°N)	Irrigated perimeter	Transfer — far	
KOGA	Ethiopia	Sub-humid highland (Amhara)	Reservoir-fed scheme, medium fields	Transfer — near	
KMW	Kenya	Equatorial mid- altitude (~0°)	Mixed irrigation, fragmented fields	Transfer — intermediate	

Seven training sites supply the L3 ground truth the models learn from. **Baixo Limpopo** and **Lamego** in Mozambique represent sub-humid to semi-arid Southern African irrigation with comparatively small and heterogeneous fields. **Awash/AWA** in the Ethiopian Rift Valley contributes hot, dry, intensively irrigated agriculture under strong evaporative demand. **Gezira/GEZ** in Sudan represents a large-field cotton and cereal system, while **Office du Niger/ODN** in Mali provides a Sahelian rice-dominated scheme with strong water-management controls. The two North African sites, **LOU** in Morocco and **JEN** in Tunisia, add Mediterranean seasonality and crop diversity. Together these seven sites were selected to make the training distribution deliberately broad rather than optimized for one biome or field pattern.

Three African zero-shot transfer sites are held entirely out of training. **MIT** in Algeria is a Mediterranean far-transfer case, geographically distant from most training sites but climatically related to the North African training examples. **KOGA** in Ethiopia is a sub-Saharan highland near-transfer case, useful for testing whether the model generalizes within East Africa under different elevation and irrigation conditions.

KMW in Kenya is an equatorial mid-altitude intermediate-transfer case, selected to probe performance under a different seasonal and field-geometry regime.

The sites range from approximately 36 degrees N in Mediterranean Tunisia, Algeria, and Morocco to approximately 25 degrees S in Limpopo, and span hyper-arid Sahelian, Mediterranean, equatorial highland, and humid-to-semi-arid tropical regimes. A model that performs consistently across this set is therefore climate-transferable across African irrigation environments rather than tuned to one local agro-ecological zone.

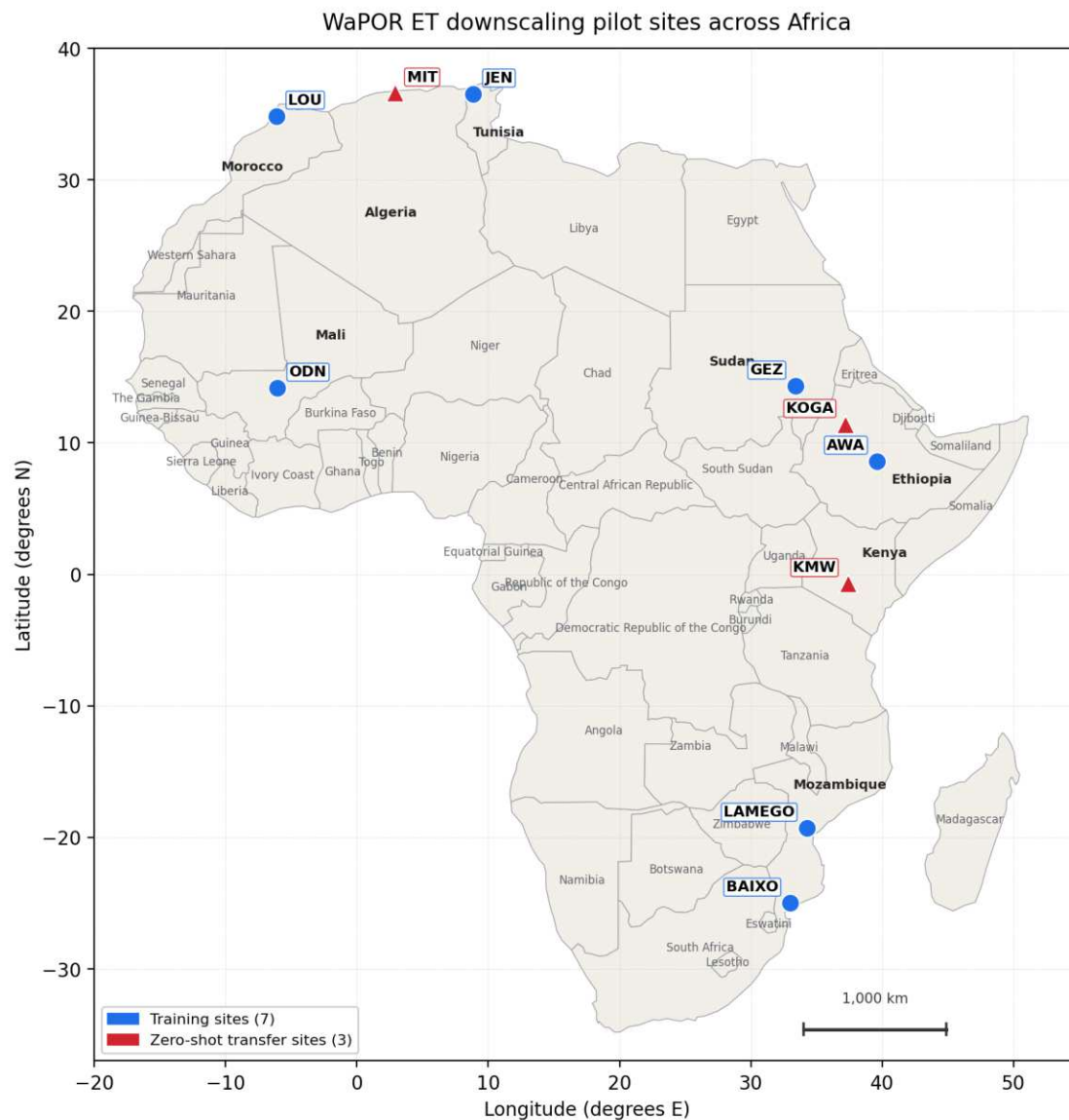


Figure 2. WaPOR ET downscaling pilot sites across Africa. Blue circles = the seven training sites; red triangles = the three African zero-shot transfer sites.

2.3 Data acquisition

All datasets are public and were acquired programmatically at dekadal (10-day) resolution and co-registered to the WaPOR Level-3 20 m grid; dekads missing any required layer (Sentinel-2 composite, L1 predictor, or L3 target) were dropped. Two acquisition channels were used: the **FAO WaPOR v3 object store for the actual-evapotranspiration-and-interception (AETI) products (FAO 2018)**, and **Google Earth Engine (GEE)** for the optical (Sentinel-2; Drusch et al. 2012), meteorological (ERA5-Land; Muñoz-Sabater et al. 2021), rainfall (CHIRPS; Funk et al. 2015), terrain (SRTM; Farr et al. 2007) and land-cover (ESA WorldCover; Zanaga et al. 2022) layers. Table 2 lists each dataset with its source, access route and role.

Table 2. *Input datasets: variables, source/collection, access route, and native resolution.*

Dataset	Variables	Source / collection	Access	Native resolution
WaPOR L3 AETI (target)	20 m ETa	L3-AETI-D mapset, FAO WaPOR v3	S3 bucket fao-gismgr-wapor-3-data (download_wap_or_l3.py)	20 m
WaPOR L1 AETI (predictor/anchor)	~300 m ETa	L1-AETI-D mapset	S3 bucket (download_wap_or_l1.py)	~300 m
Sentinel-2 L2A	B4, B8, B11 (→NDVI/NDMI/FVC)	COPERNICUS/S2_SR_HARMONIZED	GEE (fetch_s2_gee.py)	10–20 m
ERA5-Land meteorology	T2m, dewpoint, u/v wind, net solar + thermal radiation, surface pressure	ECMWF/ERA5_LAND/DAILY_AGGR	GEE (data_prep_scripts)	~9 km
CHIRPS rainfall	10-day + lagged accumulations	UCSB-CHG/CHIRPS/DAILY	GEE	~5 km
SRTM terrain	DEM, slope, aspect	USGS/SRTMGL1_003	GEE	30 m
ESA WorldCover	land cover (urban/water mask)	ESA/WorldCover/v200	GEE	10 m

The ET products (L1 and L3 AETI-Dekadal mapsets) were pulled per dekad from the FAO WaPOR v3 public bucket. Sentinel-2 was composited in GEE as described in §2.4.1; ERA5-Land was reduced to a dekadal mean and is used only to compute the FAO-56 Penman–Monteith (PM) PET ceiling for PRSN-PM (§2.4.2); CHIRPS, SRTM and WorldCover were aggregated/resampled to the 20 m grid (continuous by

bilinear interpolation, categorical by nearest-neighbour). Service-account GEE authentication and the full per-step pipeline (https://github.com/iwmihq/wapor_africa) make the acquisition reproducible for any new WaPOR L3 site.

2.4 Model inputs

All models receive the same nine input channels at 20 m (Table 3), so that any performance difference is attributable to the model, not to the information available to it. These channels were established in our prior WaPOR ETa downscaling work, where feature-subset optimisation identified them as the most important predictors (Kiala et al. 2026).

Table 3. *The nine 20 m model input channels, identical for every model.*

#	Channel	Description
0–2	B4, B8, B11	Sentinel-2 red, NIR, SWIR ($\div 10000$)
3	ETa300m	WaPOR coarse ET reprojected to 20 m ($\div 100$) — the L1 anchor
4	NDVI	$(B8 - B4) / (B8 + B4)$
5	NDMI	$(B8 - B11) / (B8 + B11)$
6	FVC	fractional vegetation cover, clipped [0,1]
7–8	$\sin(\text{DOY}), \cos(\text{DOY})$	seasonal phase

PRSN additionally reads terrain and antecedent-rainfall bands for its physics terms only. These are not fed to the learned backbone, so the predictive feature set is identical across all models.

2.4.1 Sentinel-2 preprocessing

The optical predictors derive from **Sentinel-2 Level-2A surface reflectance** (COPERNICUS/S2_SR_HARMONIZED, bottom-of-atmosphere, harmonised across the pre-/post-2022 processing-baseline offset; Drusch et al. 2012). For each WaPOR dekad we build a cloud-masked composite over the site AOI:

1. **Scene selection.** All S2 scenes intersecting the AOI within the dekad window are retrieved. The window is widened slightly beyond the nominal 10 days where needed to recover clear pixels. Scenes are pre-filtered on `CLOUDY_PIXEL_PERCENTAGE` < 60 %.
2. **Per-pixel cloud/shadow masking.** The scene-classification layer (SCL) is used to keep only valid surface classes: vegetation (4), bare soil (5), water (6), unclassified (7), and snow/ice (11). Cloud, cloud shadow, cirrus, and saturated pixels are discarded. Where the scene-level cloud filter leaves a sub-tile empty, it is recomputed without the scene filter, relying solely on the SCL pixel mask.

3. **Compositing.** The masked stack is reduced by a **per-pixel median** over the dekad, which suppresses residual haze and co-registration noise. Bands **B4 (red)**, **B8 (NIR)**, **B11 (SWIR)** are retained as reflectance $\times 10\,000$.
4. **Gridding.** Tiles are mosaicked, clipped to the AOI, and resampled to the common **20 m** grid co-registered with the WaPOR L3 target. Reflectances are rescaled to $[0,1]$ ($\div 10\,000$). The vegetation/moisture indices are then derived on this grid: $\text{NDVI} = (\text{B8} - \text{B4}) / (\text{B8} + \text{B4})$, $\text{NDMI} = (\text{B8} - \text{B11}) / (\text{B8} + \text{B11})$, and FVC (clipped to $[0,1]$). Day-of-year is encoded as sin/cos to give the seasonal phase channels.

Identical preprocessing is applied at every site, including the zero-shot transfer sites, so cross-site differences reflect surface and climate, not pipeline changes.

2.4.2 Ancillary inputs for the physics-based model

PRSN's learned backbone receives exactly the same nine channels as the direct models. Its physical constraints additionally consume geography-blind ancillary fields that are continentally available, so the model stays deployable beyond the L3 footprint. These fields are listed in Table 4:

Table 4. *Ancillary fields consumed by PRSN's physics terms only (never by the predictive backbone).*

Input	Source	Used by
Slope, aspect (sin/cos), elevation	SRTM (USGS/SRTMGL1_003), 20 m	mass-redistribution texture + energy/PET topographic modifier
Antecedent rainfall (10-day + lagged 10-day)	CHIRPS daily, accumulated	advection gate; Budyko aridity supply term
FAO-56 Penman–Monteith reference ET (PET)	ERA5-Land dekadal-mean meteorology (2 m air & dewpoint temperature, 10 m wind, net solar + net thermal radiation, surface pressure)	thermodynamic-anchor soft- hinge ceiling

The PET ceiling is computed as FAO-56 PM reference ET from ERA5. It is bilinearly downscaled to 20 m and sharpened by an SRTM clear-sky cos-incidence modifier (mean-preserving), then used as a soft upper bound on predicted ET. These ancillary fields enter only the physics terms and the loss, never the predictive backbone, so the head-to-head accuracy comparison against the direct models remains on an identical predictive feature set.

Note. Two energy-layer variants are evaluated. The default PRSN uses a *relative* energy-proxy ceiling ($\cos(\text{slope}) \cdot (1 - \text{albedo})$, block-normalised). A second variant, **PRSN-PM**, replaces it with the full **FAO-56 Penman–Monteith PET** thermodynamic anchor described above (ERA5 + SRTM, soft-hinge), faithful to the architecture specification. The head-to-head of both against the direct models is reported in §3.4.

2.5 Training and evaluation protocol

All seven training sites are used to train every model for 100 epochs. Only the three multi-year sites yield an unseen hold-out year. The four single-year sites contribute agro-climatic diversity to training but no held-out year, and hence no reportable in-domain skill, which is why they appear in no results table. Evaluation is therefore reported at six sites, split into:

- **In-domain** hold-out years at the three multi-year training sites: Baixo (2025), Lamego (2022), and Tunisia (JEN) (2025), reported per site (Table 6). An earlier pixel-count-weighted Baixo+Lamego combination has been superseded by this per-site breakdown.
- **Zero-shot transfer** to three sites never seen in training: MIT (Algeria, Mediterranean), KOGA (Ethiopia, sub-Saharan highland), and KMW (Kenya, mid-altitude).

Training details. Models are trained on random 256×256 patches with the masked Huber loss ($\delta = 5$), AdamW, and a OneCycle learning-rate schedule (see Appendix B for per-model configuration and cost). All models use identical on-the-fly **data augmentation** (Shorten and Khoshgoftaar 2019): random horizontal and vertical flips and 90° rotations, i.e. the eight symmetries of the D4 dihedral group, applied jointly to the input channels and the target. These transforms are label-preserving for this reflection- and rotation-agnostic downscaling task. They are applied identically across every model, so they do not confound the model comparison. No photometric (radiometric) or geometric-scaling augmentation is used, so the fixed L1→L3 spatial-scale relationship is never distorted. Augmentation is applied only during training; validation and hold-out evaluation use unaugmented patches.

Evaluation metrics. Let y_i be the WaPOR L3 ETa at valid pixel i , $\hat{y}_i$ the prediction, $\bar{y}$ the L3 mean, and N the number of valid pixels. Performance is quantified with five complementary metrics (Eqs. 2–3), all computed against the L3 ETLook label over the CatBoost-fair (CB-fair) valid-pixel mask, i.e. the same valid-pixel mask used to score the operational CatBoost baseline (Kiala et al. 2026), so that all models are compared on identical pixels:

$$\text{RMSE} = \sqrt{1/N \sum_i (\hat{y}_i - y_i)^2}, \quad \text{MAE} = 1/N \sum_i |\hat{y}_i - y_i|, \quad \text{Bias} = 1/N \sum_i (\hat{y}_i - y_i), \quad (2)$$

$$\text{RRMSE} = \text{RMSE}/\bar{y} \times 100\%, \quad R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2}. \quad (3)$$

Root-mean-square error (RMSE) weights large deviations; mean absolute error (MAE) is the robust mean deviation; bias is the signed systematic error (positive = overestimation); relative RMSE (RRMSE) normalises RMSE by mean ET to compare across sites/seasons of differing magnitude; R^2 is the explained variance. Metrics are computed **per dekad and per site** (aggregated to pixel-count-weighted site and combined values). All values below are in **mm dekad⁻¹**; division by 10 gives mm day⁻¹.

Statistical uncertainty. Uncertainty in per-dekad metrics is quantified by non-parametric bootstrap resampling across dekads ($B = 5000$; 95% CI = 2.5th–97.5th percentiles), which avoids normality assumptions. Differences between models are tested on matched (site, dekad) per-dekad RMSE with the paired Wilcoxon signed-rank test (a paired non-parametric test robust to skew). Multi-scale mass consistency (§3.3) is assessed by aggregating 20 m predictions to the native ~ 300 m L1 grid and comparing to L1, against the irreducible L3-vs-L1 control.

2.6 Models compared

We compare three direct super-resolution networks and the physics-compliant PRSN (Figure 3; PRSN in Figure 1); Table 5 summarises the four models and their constraint regimes.

Why these two backbones. Rather than survey many similar architectures, we deliberately bracket the design space of learned downscalers with two deeply contrasting models. Prior ET-downscaling is dominated by per-pixel tree-based regressors (random forests, gradient boosting), which cannot exploit spatial context (Singh et al. 2014; Ke et al. 2016). The operational WaPOR baseline we compare against is one such gradient-boosted model (Kiala et al. 2026). The field also relies on conventional convolutional networks (Che et al. 2022). Within end-to-end super-resolution, purely convolutional architectures have been consistently overtaken on standard benchmarks by window-attention transformers (SRCNN, EDSR, RCAN; Dong et al. 2015; Lim et al. 2017; Zhang et al. 2018). SwinIR is the canonical, compact representative of these transformers. We therefore adopt it as the strongest small, task-specific model. We prefer it to a generic encoder–decoder (U-Net; Ronneberger et al. 2015), whose plain-convolutional capacity it subsumes. We deliberately exclude GAN-based (e.g. ESRGAN; Wang et al. 2018) and diffusion super-resolution. Both optimise *perceptual realism* and can hallucinate plausible-but-unfaithful texture. This is disqualifying when the target is a conserved physical flux rather than a visually pleasing image. At the opposite pole sits **Prithvi-V1**, the leading open geospatial *foundation* model. We chose it to test whether large-scale self-supervised pre-training beats a small specialist here. The two thus span the practical choices an operational team faces: a compact model trained from scratch versus a large pre-trained ViT. Their 50/50 ensemble tests whether the two are complementary.

SwinIR (Liang et al. 2021) is a compact Swin-transformer image-restoration network with three stages (Figure 3a). A *shallow feature extractor* (a single convolution) preserves low-frequency content. A *deep feature extractor* of four residual Swin-transformer blocks (RSTBs) then refines them, each block stacking several window-based self-attention layers (with shifted windows for cross-window interaction) followed by a convolution and a residual connection. A *reconstruction head* fuses the shallow (skip) and deep features to regress the 20 m field. We use a lightweight configuration (embed dim 96, 4·4·4·4 depths/heads, window 16; ≈ 1.8 M parameters), trained from scratch on the 9-channel stack, which is well suited to the modest size of the L3 training archive.

Prithvi-V1 (Jakubik et al. 2023) is the IBM/NASA geospatial foundation model, a temporal Vision Transformer pre-trained with a masked-autoencoder objective on >1 TB of Harmonised Landsat–Sentinel-2 imagery (Figure 3b). Inputs are patch-embedded with spatial and temporal attention. The self-supervised ViT encoder provides general-purpose geospatial features, onto which we attach a regression decoder for ET. Following the standard transfer recipe, the ~ 300 M-parameter backbone is frozen up to layer 22 and only the head (plus the final blocks) is fine-tuned. It represents the "large pre-trained foundation model" end of the design space, contrasting with SwinIR's small from-scratch model. Prithvi-V1 has previously

been used for segmentation, crop mapping, and flood detection; here we test whether its large-scale geospatial pre-training transfers to a regression-downscaling task.

Ensemble is the per-pixel 50/50 mean of the SwinIR and Prithvi-V1 predictions: a plain arithmetic average with no meta-training, stacking, or learned weighting. It is a standard variance-reducing combination of two complementary backbones.

PRSN (Physics-Reconciled Scaling Network) is the physics-compliant model detailed in §2.1 and Figure 1. A SwinIR-style texture backbone produces the weighting map W , and a separate head produces the bias correction δ . The synthesis layer $ET_{20} = W \cdot R$ (with $R = \text{softplus}(\bar{L}_1 + \delta)$) enforces exact self-closure to the reconciled coarse total R , and non-negativity. Optional energy-ceiling, advection-gate, Budyko-regime, and Penman–Monteith-anchor terms add physical realism.

ETa super-resolution with WaPOR data: SwinIR vs Prithvi-V1

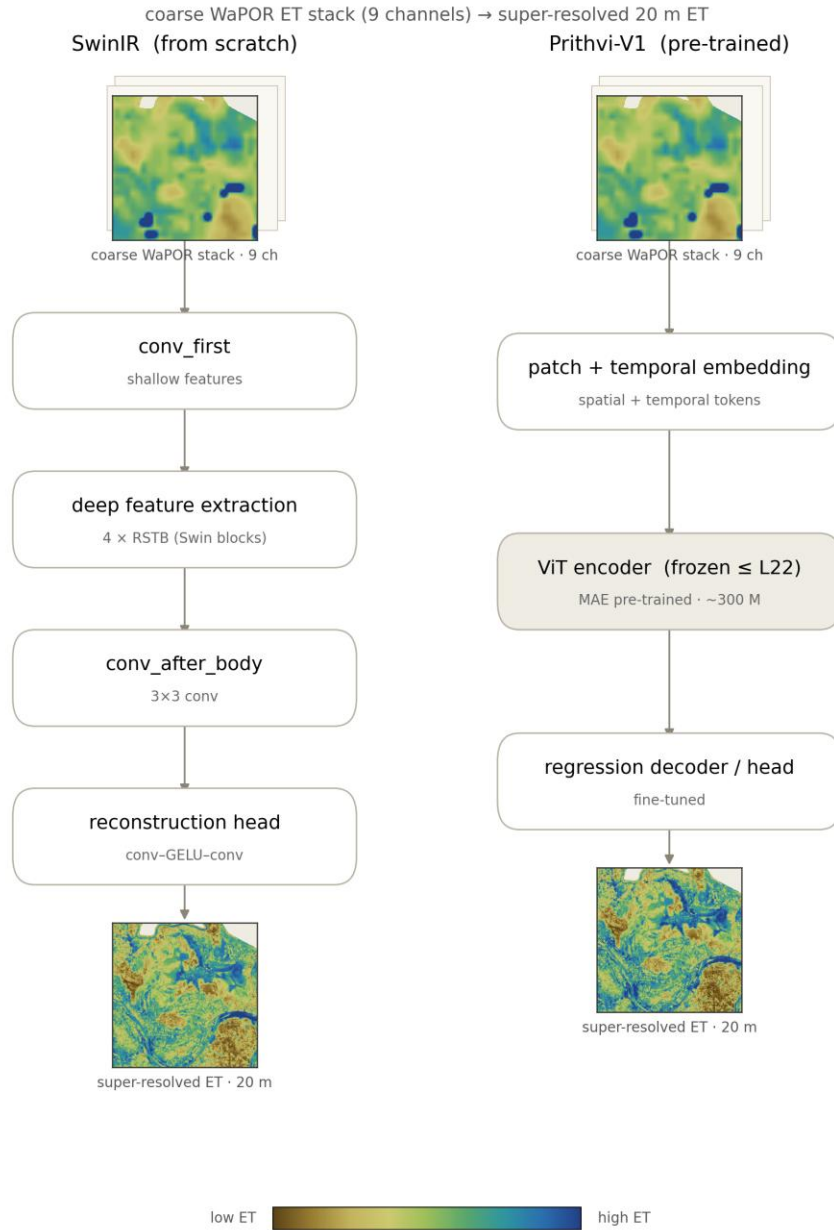


Figure 3. The two direct super-resolution models, mapping the same nine-channel coarse WaPOR ET stack to a 20 m ET field. **(left)** SwinIR, trained from scratch (Liang et al. 2021). **(right)** Prithvi-V1, the pre-trained foundation model with a fine-tuned regression head (Jakubik et al. 2023). Top thumbnails show the shared blocky ~300 m L1 input patch; bottom thumbnails show each model's 20 m output (Baixo, 2025-06-11), on the shared ET colour scale.

Table 5. Models compared and their constraint regimes (SR = super-resolution).

Family	Model	Constraint regime
Direct SR	SwinIR-L1 (~1.8 M params)	none (free regression)
Direct SR	Prithvi-V1-L1 (IBM/NASA ViT)	none
Direct SR	Ensemble (50/50 SwinIR+Prithvi)	none
Physics	PRSN	hard mass closure + $ET \geq 0$ + energy ceiling + advection gate
Physics	PRSN + Budyko	PRSN + aridity-regime attention

3. Results

3.1 Accuracy by site

Under identical inputs, sites, and training, the direct models are more accurate than the physics-compliant models at every site, both in-domain and under zero-shot transfer (per-site RMSE in Table 6). Figure 6 shows the corresponding downscaled maps for the best hold-out dekad, alongside the native 300 m input and the 20 m reference.

SwinIR (fine-tuned) is the 7-site SwinIR model, further trained for a short time on the target site's own L3 dekads. This light, site-specific adaptation makes it the single most accurate model on the genuine transfer sites (those absent from training), where it beats even the ensemble: RMSE falls by 12 % at MIT, 30 % at KOGA, and 22 % at KMW relative to the zero-shot SwinIR. The gain shows that, wherever a site's own L3 data are available, local adaptation is the largest single lever on accuracy.

†**Tunisia (JEN)** is one of the seven training sites, so it has no zero-shot setting. The value is the model's skill on a held-out future year (2025) never seen in training, the only true hold-out available for a training site. Here, fine-tuning on JEN's own ≤ 2024 data degrades the 2025 hold-out ($3.00 \rightarrow 3.53$): for a site already in the training mix, narrow per-site fine-tuning trades away the multi-site diversity that aids generalisation. The same in-domain pattern holds for Baixo and Lamego, the other two training sites evaluated on unseen hold-out years (Baixo 2025, Lamego 2022). Fine-tuning barely moves Baixo ($3.91 \rightarrow 3.77$, -3.7%) and slightly worsens Lamego ($4.10 \rightarrow 4.28$, $+4.4\%$); in neither case does it match the large gains seen on the genuine transfer sites. In short, fine-tuning helps genuine transfer sites, not in-domain ones. The full Tunisia row reproduces the paper's central ranking on an independent future-year hold-out: the direct models (SwinIR 3.00, Ensemble 3.01, Prithvi 3.23) beat the physics-compliant models (PRSN 3.58, PRSN+Budyko 3.62), and Budyko again fails to help.

Table 6. Accuracy by site: RMSE (mm dekad⁻¹). The first five model columns are the zero-shot comparison (no target-site labels) and are the paper’s core result. The final column, SwinIR (fine-tuned), is the operational extension (§4.4); it uses target-site L3 and is shown for reference only, not as part of the controlled comparison

Site	Setting	Ensemble	SwinIR	Prithvi	PRSN	PRSN+Bu dyko	SwinIR (fine- tuned)
Baixo	in-domain	3.82	3.91	3.99	4.93	5.07	3.77
Lamego	in-domain	4.11	4.10	4.24	4.33	4.44	4.28
Tunisia (JEN)	in- domain†	3.01	3.00	3.23	3.58	3.62	3.53
MIT (Algeria)	transfer	3.47	3.56	3.66	5.09	4.95	3.14
KOGA (Ethiopia)	transfer	4.06	4.35	4.21	5.36	5.81	3.06
KMW (Kenya)	transfer	4.33	4.57	4.37	5.64	5.82	3.58

(mm day⁻¹ = above ÷ 10. Bold = best zero-shot model per row; the fine-tuned column is reference only.)

Comparison against operational baselines. To place the deep models against what an operational team would otherwise use, Table 6a adds two references at the same sites and hold-outs. The first is naive L1 upsampling: the coarse ~300 m L1 ET resampled to the 20 m grid with no learning, i.e. the zero-skill floor. The second is the operational per-pixel CatBoost downscaler currently used at IWMI (Kiala et al. 2026), a gradient-boosted regressor trained on Baixo+Lamego, so those two sites are in-domain for it and the rest are transfer. The gap is large and consistent. The best direct model roughly halves the naive-upsampling RMSE at every site (e.g. Baixo 7.58→3.82; KOGA 9.00→4.06, -55 %). It also beats the operational CatBoost everywhere. The margin is modest in-domain (Baixo 4.15→3.82, Lamego 4.44→4.11) but decisive under transfer. There, the per-pixel CatBoost lacks spatial context and degrades to 5.6–8.0 mm dekad⁻¹, worse than naive upsampling at JEN and MIT, while the direct models stay at 3.0–4.3. This is the practical case for deep, spatially-aware downscaling: the value is small where a calibrated local model already exists, but large in exactly the transfer regime that motivates continental deployment. Note that the CatBoost gap conflates two advantages, spatial context and the seven-site training set, so it is an operational comparison, not a controlled-architecture one. The controlled contrast is the direct-vs-physics comparison above, where inputs and training are held identical.

Table 6a. Accuracy against baselines: RMSE (mm dekad⁻¹) at each site's hold-out. Naive L1 = coarse L1 resampled to 20 m (no learning). CatBoost = operational per-pixel downscaler (Kiala et al. 2026), trained on Baixo+Lamego, so it is in-domain there and transfer elsewhere. Best per row in bold.

Site	Setting	Naive L1 → 20 m	Operational CatBoost	Best direct (Ensemble)	PRSN-PM (physics)
Baixo	in-domain	7.58	4.15	3.82	5.32
Lamego	in-domain	5.86	4.44	4.11	4.40
Tunisia (JEN)	in-domain†	5.81	7.78	3.01	3.58
MIT (Algeria)	transfer	6.50	8.03	3.47	4.50
KOGA (Ethiopia)	transfer	9.00	6.85	4.06	5.34
KMW (Kenya)	transfer	7.30	5.62	4.33	5.05

(CatBoost is in-domain only at Baixo and Lamego; JEN/MIT/KOGA/KMW are transfer for it. Even the physics-compliant PRSN-PM — the weakest of our learned models — beats naive upsampling at every site and beats the operational CatBoost at every transfer site.)

3.2 Variance explained (R^2)

The same ranking holds for explained variance: among the zero-shot models, the direct models achieve the highest R^2 at every site (Table 7). Their explained variance is strong throughout, from 0.869–0.910 in-domain to 0.779–0.844 on the transfer sites, with the Ensemble best at four sites and SwinIR at Tunisia (and tied at Lamego). The physics-compliant models trail everywhere, and the gap widens under transfer: PRSN and PRSN+Budyko fall to 0.70–0.75 at MIT and KOGA and to just 0.64 and 0.61 at KMW, against 0.80 for the Ensemble. In other words, enforcing physics costs ≈ 0.01 –0.19 in R^2 , and most of that loss lands on the transfer sites, exactly where the coarse L1 anchor is least reliable.

Table 7. Variance explained (R^2) by site. Bold = best zero-shot model per row; SwinIR (fine-tuned) is reference only (§4.4).

Site	SwinIR	Prithvi	Ensemble	PRSN	PRSN+Bud yko	SwinIR (fine-tuned)
Baixo	0.873	0.867	0.879	0.789	0.776	0.882
Lamego	0.869	0.860	0.869	0.855	0.787	0.858
Tunisia (JEN, 2025)	0.910	0.896	0.909	0.872	0.869	0.876
MIT	0.831	0.821	0.839	0.703	0.720	0.868
KOGA	0.821	0.832	0.844	0.754	0.711	0.911
KMW	0.779	0.798	0.800	0.638	0.613	0.863

Complete per-site metrics (RMSE, MAE, RRMSE, R^2) for all five models are given in **Appendix A**.

3.2.1 Per-dekad (seasonal) accuracy

Aggregated metrics can hide seasonal structure, so we also resolve error per dekad. Figure 4 plots per-dekad RRMSE against day-of-year for each site, with evaluation years aligned by season. RRMSE is per-dekad RMSE normalised by each site's mean ET (§2.5), so error is comparable across sites of differing ET magnitude.

Three patterns emerge. First, seasonality is strong: error rises during the high-ET growing and irrigation season (mid-year at the Northern-Hemisphere Mediterranean sites; the wet season at the Southern sites) and falls in the low-ET dormant period, tracking the seasonal ET cycle. Second, the ordering is persistent: the direct model is below both physics models in the large majority of individual dekads, not merely in the seasonal mean, which is what the paired per-dekad tests of §3.5 formalise. Third, the PM anchor's transfer benefit is concentrated in the high-error peak-season dekads at MIT and KMW, where it pulls PRSN-PM toward, but not past, the direct curve.



Figure 4. Per-dekad RRMSE (%) versus day-of-year by site, for the direct SwinIR and the two physics models. Error peaks in the high-ET growing season; the direct model lies below both physics models in nearly every dekad.

3.3 Physical-consistency diagnostics

Beyond aggregate accuracy, we ask what each model family delivers in physical consistency, and at what cost to fidelity. Figure 5 gathers four diagnostics on the Baixo hold-out. Panels (a, b) compare each model to the L3 reference pixel-by-pixel: the direct SwinIR tracks the 1:1 line tightly, whereas the physics-compliant PRSN shows a broad over-prediction cloud above it. Panels (c, d) show the two consistency properties the physics family is constructed to guarantee. In (c), both SwinIR (RMSE 5.36) and PRSN-PM (3.03) re-aggregate to within the irreducible L3→L1 product floor (6.00), and PRSN-PM is the tighter of

the two; PRSN's exact machine-precision closure ($\approx 3 \times 10^{-6}$) is to its own reconciled coarse total R, not to observed L1, a guarantee the direct model has no analogue of. In (d), non-negative flux ($ET \geq 0$) is guaranteed by construction for PRSN but not for the direct models. Together the panels make the paper's central point: the physics family's advantage is consistency, not fidelity.

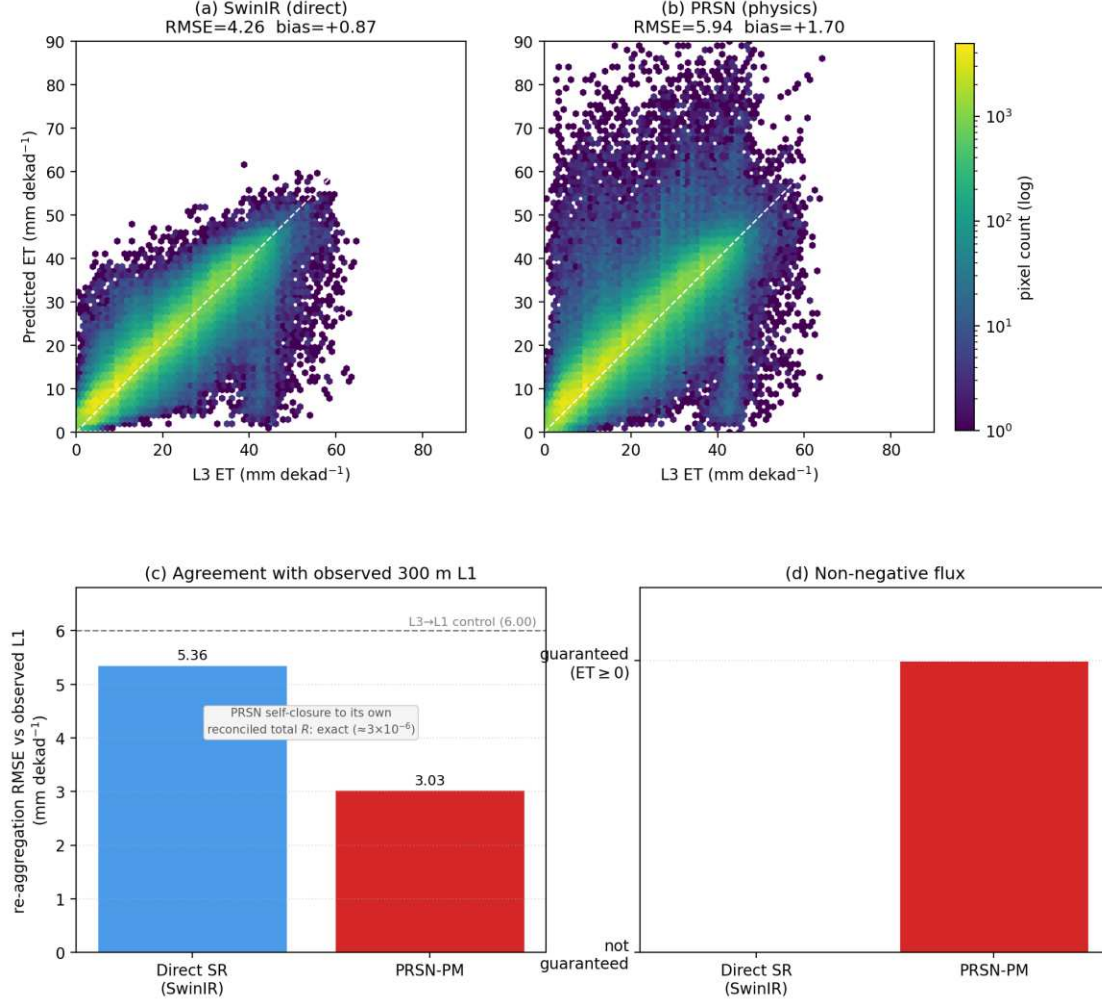


Figure 5. Physical-consistency diagnostics on the Baixo in-domain hold-out (2025). **(a, b)** Predicted-vs-L3 ET density (log-count hexbins, four subsampled dekads); dashed line = 1:1. Panel RMSE and bias are computed on the dekads shown and differ slightly from Table 6. **(c)** Re-aggregation RMSE against the observed 300 m L1 (mm dekad⁻¹; Table 8); dashed line = the irreducible L3→L1 floor. **(d)** Non-negativity guarantee by model family.

The diagnostics above state the consistency contract qualitatively; here we quantify them. We aggregate each model's 20 m field (area-weighted) to the native ~300 m WaPOR L1 grid and compare to L1. Because L1 and L3 are independently produced products, even the L3 reference does not aggregate exactly to L1 — that L3-vs-L1 discrepancy is the irreducible floor against which a model's closure should be judged (Peng

et al. 2017; Yang et al. 2020). Table 8 reports the aggregation bias, RMSE and non-negativity for each model against this control.

Table 8. Multi-scale mass conservation: 20 m predictions aggregated to the ~ 300 m L1 grid vs L1.

Aggregated $\rightarrow$ observed 300 m L1 (in- domain hold-out)	Bias vs L1	RMSE vs L1	% pixels < 0
Direct family (SwinIR)	−0.53 (−2.7 %)	5.36	0.00004 %
Physics family (PRSN- PM)	−0.22 (−1.2 %)	3.03	0
L3 reference $\rightarrow$ L1 (irreducible control)	−0.66 (−3.1 %)	6.00	—

(Pooled over the in-domain hold-out, Baixo 2025 + Lamego 2022. Each model family is represented by its strongest member — SwinIR for the direct models, PRSN-PM for the physics models. All rows aggregate the 20 m field, area-weighted, to the native ~ 300 m L1 grid, scoring cells with ≥ 90 % valid coverage. PRSN-PM’s **exact** machine-precision closure is to its own reconciled coarse total R , not to observed L1; the row reports its **approximate** agreement with observed L1.)

Three points:

1. **The physics guarantees are real, and the direct model already approximates them.** PRSN guarantees, by construction, that the 20 m field re-aggregates *exactly* to its predicted coarse total R (machine-precision residual $\approx 3 \times 10^{-6}$) and is strictly non-negative. The *unconstrained* direct model is not exact, but it already aggregates to within −2.7 % of L1, inside the L3-vs-L1 product discrepancy (−3.1 %), and is essentially non-negative already (3.8×10^{-5} % of pixels below zero). So the direct model approximates both physical properties without being told to; the physics model’s contribution is to make them *exact and hard* (point 2 quantifies the coarse-scale gap).
2. **The constraints also buy tighter coarse-scale agreement, at the cost of fine-scale fidelity.** Because δ shifts the block total, PRSN’s exact self-closure to R does not, in principle, guarantee agreement with the observed L1 (the quantity a closed water-accounting budget actually requires). Measured directly, PRSN-PM’s aggregate sits **−1.2 % from observed L1 (RMSE 3.03), approximate rather than exact**, yet in fact *closer* to L1 than the direct model (−2.7 %, 5.36), because the learned δ keeps R near the L1 block mean. So the trade-off is scale-dependent and explicit: **enforcing physics buys exact self-consistency and tighter agreement with the coarse 300 m L1, but this comes at the price of the fine-scale (20 m) fidelity that is the actual downscaling objective** (§3.1). The direct model spends its capacity fitting the 20 m target; PRSN spends part of it honouring the coarse anchor. Which matters depends on the use case.
3. **Where the hard guarantee still matters.** Exact closure to R is a guarantee for *every* cell, not a statistical tendency. It is required when the 20 m field must be coupled into a closed budget against the model’s own coarse total. That, together with the tighter coarse-scale agreement above, is the operational niche for physics-compliance.

This is the empirical counterpart to the paper’s thesis: hard constraints deliver exact per-cell self-consistency and tighter coarse-scale agreement, but they trade away fine-scale accuracy (§3.1, §3.4), so they are the right choice only when coarse-scale conservation, not 20 m fidelity, is the priority.

3.4 Proxy versus Penman–Monteith energy anchor

A natural objection to the result above is that PRSN’s energy ceiling is only a relative proxy ($\cos(\text{slope}) \cdot (1 - \text{albedo})$), so its underperformance might reflect a weak physics implementation rather than a limitation of physics-anchoring itself. We therefore trained a second 7-site PRSN, **PRSN-PM**, whose thermodynamic anchor is a full **FAO-56 Penman–Monteith reference ET** (Allen et al. 1998) computed from ERA5 meteorology and SRTM terrain (§2.4.2), as a soft-hinge ceiling, the strongest, most physically faithful version of the constraint. Mass closure remained machine-precision ($\approx 3 \times 10^{-6}$) throughout. Table 9 compares the proxy and Penman–Monteith anchors against the best direct model at each site.

Table 9. Proxy vs FAO-56 Penman–Monteith energy anchor: RMSE by site against the best direct model. The PM anchor *narrows the physics-vs-direct gap on transfer but does not close it: the best direct model remains most accurate at every site.*

Site	Setting	Best direct	PRSN (proxy)	PRSN+Budyko	PRSN-PM
Baixo	in-domain	3.82	4.93	5.07	5.32
Lamego	in-domain	4.10	4.33	4.44	4.40
MIT (Algeria)	transfer	3.47	5.09	4.95	4.50
KOGA (Ethiopia)	transfer	4.06	5.36	5.81	5.34
KMW (Kenya)	transfer	4.33	5.64	5.82	5.05
JEN (Tunisia, 2025)	in-domain†	3.00	3.58	3.62	3.58

(RMSE, mm dekad⁻¹; bold = best per row. “Best direct” = the lowest of Ensemble/SwinIR.)

Two findings:

1. **The real PM anchor does improve transfer over the proxy: the held-out transfer average falls from 5.36 (proxy) to 4.96 (PRSN-PM), −7.5 %, with the largest gains at MIT (5.09→4.50) and KMW (5.64→5.05).** A physically-grounded energy ceiling is therefore a genuinely better out-of-domain anchor than the geometric proxy, confirming the proxy *was* the weaker implementation. (In-domain it is marginally worse (Baixo 4.93→5.32), where the extra constraint has no domain gap to bridge.)
2. **It still does not close the gap to the direct models.** PRSN-PM remains less accurate than the best direct model at every site (e.g. MIT 4.50 vs 3.47; KMW 5.05 vs 4.33; in-domain Baixo 5.32 vs 3.82). The strongest, fully physically-faithful energy layer narrows the physics penalty on transfer but does not eliminate it.

This is the decisive control for the paper’s thesis: even when the physics constraint is implemented at its strongest, exact mass closure plus a true Penman–Monteith energy anchor, hard physical constraints buy *consistency and somewhat better transfer robustness*, but **not** competitive accuracy. Direct super-resolution remains the more accurate choice.

3.5 Statistical significance

To confirm the model differences are not sampling noise, we bootstrap the mean per-dekad RMSE ($B = 5000$) and apply paired Wilcoxon signed-rank tests on matched (site, dekad) RMSE (§2.5). Per-dekad means differ slightly from the pixel-count-weighted aggregates of §3.1, since dekads are weighted equally here, but the ranking is identical. Table 10 reports the mean dekadal RMSE with 95 % bootstrap confidence intervals.

We flag one caveat on the inference. The bootstrap and the paired Wilcoxon test treat per-dekad values as independent, whereas adjacent dekads within a site are seasonally autocorrelated. The effective sample size is therefore smaller than the nominal $n = 48\text{--}106$, and the reported CIs and p-values are somewhat optimistic. The effect sizes here are large, however: the direct-vs-physics gaps are multiples of their CIs and hold in the large majority of individual dekads, so significance is not in doubt. A site/season block-resampling scheme would be the more conservative choice, and is the recommended refinement for borderline comparisons.

Table 10. Mean dekadal RMSE with 95 % bootstrap confidence intervals ($B = 5000$).

Setting	Direct (SwinIR)	PRSN (proxy)	PRSN-PM
In-domain	3.90 [3.55, 4.28]	4.47 [3.95, 5.01]	4.69 [4.26, 5.14]
Transfer	4.09 [3.88, 4.31]	5.26 [5.07, 5.48]	4.79 [4.70, 4.89]

(Mean dekadal RMSE, mm dekad⁻¹; 95 % bootstrap CI in brackets. Bold = lowest per row.)

Paired Wilcoxon tests:

- **Direct < PRSN-PM, in-domain:** median $\Delta\text{RMSE} -0.68$ mm dekad⁻¹, $p = 2 \times 10^{-10}$ ($n = 72$ dekads).
- **Direct < PRSN-PM, transfer:** median $\Delta\text{RMSE} -0.30$, $p = 8 \times 10^{-8}$ ($n = 106$).
- **Direct < PRSN (proxy), in-domain:** median $\Delta\text{RMSE} -0.72$, $p = 8 \times 10^{-13}$ ($n = 48$).
- **PRSN-PM < PRSN (proxy), transfer:** median $\Delta\text{RMSE} -0.17$, $p = 3 \times 10^{-5}$. The PM anchor’s transfer gain over the proxy is significant.

The direct models’ accuracy advantage over both physics variants is therefore statistically significant, in-domain and on transfer. The PM anchor’s improvement over the proxy on transfer is also significant. The effects reported above are real, not noise.

3.6 Qualitative comparison of downscaled maps

Figure 6 illustrates the results spatially across all six sites, each at its lowest-RMSE hold-out dekad. The native 300 m L1 input (left) is visibly blocky, whereas the direct models reconstruct sharp, field-scale texture matching the 20 m L3 reference (right). PRSN recovers similar structure but is smoother and

blockier at its ~ 300 m mass-closure block boundaries, with higher error at every site. This is most visible at the low-ET Mediterranean sites (Tunisia, Algeria), where its RMSE is roughly double. Figure 6 is thus the visual counterpart to the quantitative ranking: direct super-resolution both scores better and looks right.

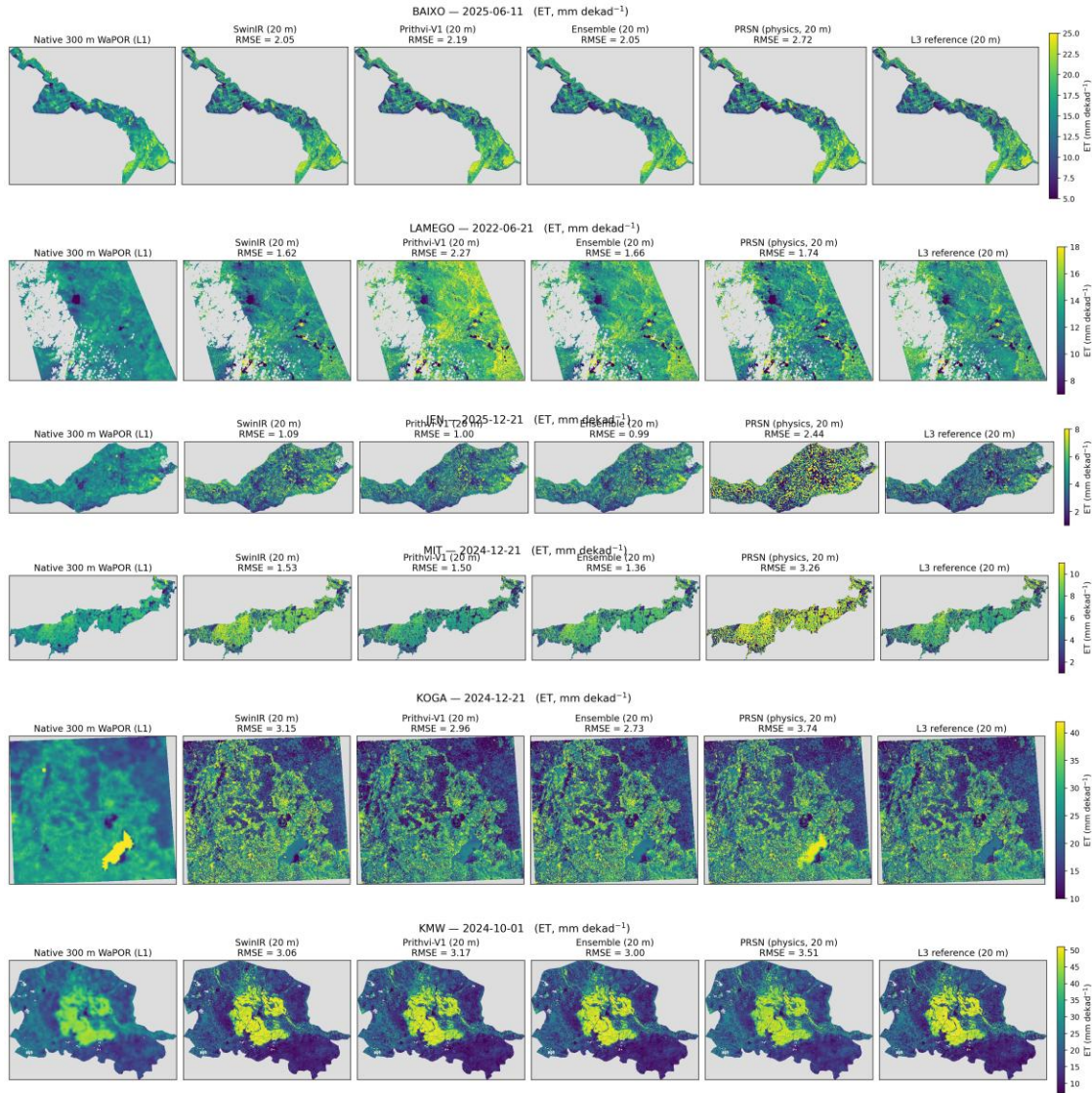


Figure 6. Downscaled ET at each site's lowest-RMSE hold-out dekad (rows: Baixo, Lamego, JEN, MIT, KOGA, KMW; per-site colour scales, mm dekad⁻¹). Columns, left to right: native 300 m L1 input; SwinIR, Prithvi-V1, and Ensemble (direct); PRSN (physics-compliant); 20 m L3 reference. Panels are annotated with RMSE against L3 on the cloud-screened valid mask.

4. Discussion

This study asked whether enforcing physical constraints improves deep-learning downscaling of WaPOR ET from 300 m to 20 m, holding inputs, training sites, and evaluation protocol identical across model families. The answer is no: the premise that hard constraints improve accuracy does not hold for this task.

The result holds both across the in-domain/transfer split above and under the strongest physics implementation we could build, PRSN-PM, whose energy term is a full FAO-56 Penman–Monteith anchor (§3.4). Even this most physically faithful constraint only narrows the transfer gap without closing it. The conclusion is therefore not an artefact of a weak energy term.

4.1 Why direct regression wins: constraints buy consistency, not accuracy

Three mechanisms reinforce one another. First, **freedom to correct the coarse bias**. L1 and L3 differ systematically, and when the learned correction δ is imperfect, hard closure propagates the residual coarse bias into the fine field. The direct model can instead overwrite the coarse value wherever the covariates warrant, which is exactly where the L1–L3 discrepancy is largest. Second, **the constraint adds no information here**. Unlike forward emulation, where a conservation law injects structure the data lack (Beucler et al. 2021; Willard et al. 2022), in this inverse refinement the coarse anchor and L3 target already embed the physics. The constraint removes degrees of freedom and slightly distorts the within-block distribution without supplying anything new. Third, **optimisation focus**. The direct model spends its entire gradient signal on the L3 error, while PRSN must also satisfy closure, δ -calibration, and the energy/Budyko terms; the Budyko mechanism compounds this by conditioning on a noisy aridity proxy. Consistent with these mechanisms, the physics penalty is largest where L1 and L3 diverge most (the transfer sites), and the free model is already approximately mass-consistent (cf. Peng et al. 2017), so the constraint reclaims little. The accuracy ceiling is thus set by the L3 target and the inputs, not by model expressivity. Physical constraints are best understood as a consistency contract, not an accuracy lever (Karpatne et al. 2017; Reichstein et al. 2019): pay for them only when downstream use requires exact re-aggregation (e.g. water-accounting budgets; Karimi et al. 2013; Blatchford et al. 2020) or guaranteed non-negativity.

4.2 Comparison with previous studies

Accuracy is competitive with prior ET-downscaling work. Our direct models (in-domain R^2 0.87–0.91, transfer R^2 0.78–0.84, RRMSE $\approx$ 17–25 %; Appendix A) compare favourably with the literature. Ke et al. (2016) reported RRMSE $\approx$ 22 % and in-situ $R^2 = 0.56$ for Random-Forest downscaling of MODIS 1 km ET to 30 m; Singh et al. (2014) reported $R^2 = 0.52$ –0.88 fusing MODIS and Landsat ET; Che et al. (2022) reported $R^2 = 0.67$ and relative RMSD = 14.25 % for a deep network sharpening MODIS 500 m ET to 30 m. Only Che et al.'s proportional error is lower than ours, and theirs is an easier setting: flux-validated 30 m sharpening at an 8-day step, versus dekadal ≈ 300 m $\rightarrow$ 20 m over fragmented African irrigation. The physics-compliant variants (RRMSE 19–35 %) fall to the weaker end of this same range, underlining that the constraint, not the data or backbone, is what costs accuracy.

Relation to the physics-informed-ML literature. The evidence that physical constraints improve accuracy and generalisation comes largely from forward emulation (Karpatne et al. 2017; Reichstein et al. 2019; Beucler et al. 2021; Willard et al. 2022); as discussed above, the mechanism reverses in our inverse setting. Our result therefore nuances, rather than contradicts, that literature: a constraint's value depends on whether it injects missing information or re-imposes what the data already contain. Our multi-scale finding,

that a well-trained unconstrained downscaler is approximately scale-consistent without enforcement (Section 3.3), likewise echoes soil-moisture and ET downscaling validation (Peng et al. 2017).

4.3 Per-site accuracy patterns

Accuracy varies systematically across the sites, and two regularities explain most of it (Appendix A).

Absolute error scales with ET magnitude. Absolute RMSE tracks each site's mean ET, lowest at low-ET JEN and highest at high-ET KMW and KOGA, but the RRMSE ranking inverts once normalised, making KOGA and KMW the most accurate in relative terms. The same scaling holds seasonally, with error peaking in the high-ET growing season (Figure 4). Between-site RMSE differences therefore largely reflect ET climatology, not differing model skill, and RRMSE is the fair cross-site metric.

Regime learnability, not geographic distance, sets the skill ceiling. The Mediterranean North-African sites are the best downscaled in explained-variance terms (JEN R^2 0.91, MIT 0.83): long clear-sky dry seasons yield clean Sentinel-2 composites, the bare-soil↔irrigated contrast maps strongly onto the optical indices (Gao 1996; Carlson & Ripley 1997), and their large geometric fields are spatially coherent at the 300 m→20 m step. The equatorial and highland sites score lower (KMW R^2 0.78, KOGA 0.82–0.84): cloudier rainy seasons degrade the composites, spatially uniform background ET weakens the vegetation–ET contrast, and fields are more fragmented (You et al. 2011; Xie et al. 2014). This is why the far transfer site MIT is downscaled better than the intermediate site KMW: transfer degrades with agro-climatic dissimilarity and intrinsic learnability, not raw distance from the training cluster.

The physics penalty is also site-dependent. PRSN's gap to the direct models is largest at the transfer sites, where the coarse L1 anchor is least reliable (MIT RRMSE $\approx$ 24 % direct vs $\approx$ 35 % proxy and $\approx$ 31 % PM; KMW $\approx$ 17 % vs $\approx$ 22 %), and smallest in-domain at Lamego ($\approx$ 18 % vs $\approx$ 19 %). This matches the propagated-coarse-bias mechanism of Section 4.1: where L1 tracks L3 well, the hard anchor costs little; where they diverge, enforcing closure hurts more.

4.4 Operational extension: where local adaptation helps

The controlled zero-shot comparison (§3.1–§3.4) is the paper's core. This subsection steps outside it, to the operational regime after target-site L3 becomes available. There, light per-site fine-tuning is the single largest accuracy lever we found, far larger than any architectural or constraint choice (Table 6). Because it consumes target-site labels, it complements rather than qualifies the zero-shot result.

It helps most on genuine transfer sites. Fine-tuning the 7-site SwinIR on a held-out site's own L3 cut RMSE by 12–30 %. The single fine-tuned SwinIR became the most accurate model overall (Table 6). The gain scales with distance from the training distribution, and is largest at KOGA, a regime under-represented in training.

It can hurt in-domain sites. On Tunisia (JEN), a training site evaluated on its unseen 2025 year, fine-tuning on the site's own $\leq$ 2024 history raised hold-out error (3.00→3.53). Narrowing an already in-

distribution model to one site's idiosyncrasies discards the multi-site diversity that regularises it against an unseen year. The operational rule is asymmetric: fine-tune for regimes the base model has not seen; keep the diverse base where the site is already in-distribution.

Fine-tune the single model, not the ensemble. The small, fully-trainable SwinIR adapts far more effectively than the frozen-backbone Prithvi ViT (fine-tuned RMSE 3.66–3.93; Jakubik et al. 2023). Blending the weakly-adapted Prithvi back in therefore dilutes the result. The ensemble that wins zero-shot loses to the single SwinIR after fine-tuning.

Fine-tuning is thus the constructive counterpart to the paper's negative finding. The accuracy headroom that physical constraints failed to unlock is unlocked instead by data-driven local adaptation, consistent with adaptation gains being largest where local data shift the input distribution (Bommasani et al. 2021). The binding constraint for this task is information about the target site, not physical admissibility.

4.5 Why the compact SwinIR outperforms the Prithvi foundation model

A striking secondary result is that the ≈ 1.8 M-parameter SwinIR, trained from scratch, matches or beats the ~ 300 M-parameter Prithvi-V1 foundation model. Zero-shot, SwinIR is the more accurate single model at four of six sites; after per-site fine-tuning it leads everywhere (Table 6). This runs against the expectation that a large pre-trained backbone should dominate. Four mechanisms explain it.

Task–pretraining mismatch. Prithvi is pre-trained by masked-autoencoder reconstruction of optical reflectance (He et al. 2022; Claverie et al. 2018). Our task is regression of a physical flux from a stack mixing reflectance with engineered covariates never seen in pre-training. The frozen features are therefore only loosely aligned with the ET–covariate mapping.

Frozen versus trainable capacity. Prithvi's backbone is frozen to layer 22, so its effective trainable capacity is a few million parameters, comparable to SwinIR, but constrained to reshape features it cannot adapt. SwinIR is trained end-to-end for this exact objective. The fine-tuning experiment makes this concrete: SwinIR's error drops 12–30 % while Prithvi stalls at 3.66–3.93.

Inductive bias for super-resolution. SwinIR was designed for image restoration (Liang et al. 2021; Liu et al. 2021): a low-frequency skip carries the coarse signal while windowed attention refines local texture, precisely what coarse-to-fine downscaling needs. Prithvi's 16×16 patch embedding coarsens spatial detail for semantic representation (Dosovitskiy et al. 2021); recovering 20 m field boundaries works against its design. The maps (Figure 6) show the consequence: SwinIR's fields are sharp, Prithvi's slightly smoother.

Data scale. Foundation-model advantages are largest in the low-label regime (Bommasani et al. 2021). Here the seven-site archive is ample for a 1.8 M-parameter network but far too small to justify re-training 300 M parameters.

The implication is not that foundation models are useless for Earth observation. Rather, for a narrow, well-supervised regression task, a small task-specific architecture is more accurate, cheaper to train (Appendix B), and far more responsive to the fine-tuning that is this problem's largest accuracy lever.

4.6 Limitations

- The reference is itself a model. Every "accuracy" figure is fidelity to WaPOR L3 ETLook, an unconstrained model (FAO 2018), not measured flux. This risks circularity: ETLook does not enforce mass conservation, so a mass-conserving model may be penalised for not reproducing L3's own non-conservative behaviour. The defensible reading is therefore "enforcing physics reduces fidelity to a non-physics reference," not "enforcing physics reduces ET accuracy." No flux station coincides with our pilots, so independent validation against eddy-covariance or lysimeter data, as in prior WaPOR evaluations (Blatchford et al. 2020), remains the outstanding step. Until then the accuracy claim is stated relative to L3, not true ET.
- Transfer is evaluated at three African sites; broader validation across irrigation systems and years remains needed (cf. Blatchford et al. 2020).
- No full constraint ablation. We compare PRSN, PRSN+Budyko, and PRSN-PM, but do not isolate each constraint's accuracy cost in turn. Mass closure and non-negativity are architecturally entangled in the block-softmax synthesis, so a term-by-term ablation would require re-architecting and retraining several variants. We leave this to future work; the PM-anchor control already shows the energy term is not the binding limitation.

4.7 Future directions

Several directions follow from these findings:

- Ground-truth validation. The priority is validation against independent in-situ flux (eddy-covariance, lysimeter) and thermal cross-checks (e.g. Landsat-based ET), to establish how well either family tracks true ET rather than the L3 model.
- Raise the ceiling, not the constraint. The larger gains lie in the inputs and target: thermal predictors carrying energy-balance information the optical stack lacks, better gap-filled composites, and bias-corrected L3 references.
- Consistency without the accuracy cost. A post-hoc conservation step applied to the direct model's output could deliver exact closure where water accounting needs it, while preserving direct-model accuracy.
- Operational transfer and uncertainty. Continental deployment where only L1 exists calls for per-pixel uncertainty estimates and a principled rule for when to fine-tune. Pretraining on the continental L1/Sentinel-2 record may further improve zero-shot transfer.
- Broader evaluation. More irrigation systems, hold-out years, and field geometries would test the learnability-not-distance hypothesis (§4.3) more stringently.

5. Conclusion

This study delivered the first controlled test of whether physical constraints improve deep-learning ET downscaling, comparing direct and physics-compliant models on identical inputs, sites, and training. We

find they do not: direct super-resolution outperforms the mass-conserving PRSN at every site, both in-domain and under transfer, and the Budyko mechanism provides no benefit. The result is stated relative to L3 deliberately. The reference is a model, not measured flux, so the direct models may partly be learning L3's errors; an independent flux check remains the outstanding validation step (§4.6). This does not blunt the practical contribution. L3 is the best field-scale reference available, so the model that reproduces it most faithfully is the more useful product now. Hard constraints still deliver exact self-closure and non-negativity, and remain the right choice where those are required. For fidelity, use direct super-resolution trained on diverse L3 sites, fine-tuned where L3 is available.

A model that learns from the few schemes with 20 m reference data and transfers to unseen agro-climates could synthesise field-scale ETa where none exists. That is a real prospect for water accounting across data-poor Africa. We stress potential, not demonstrated fact: our evaluation is confined to sites where L3 exists, so no-reference deployment remains the next step. This work provides the evidence base for it: which model family to deploy, when constraints are worth their cost, and where the accuracy ceiling lies.

Appendix A — Full per-site accuracy

Table A1 reports complete per-site metrics for all five models (RMSE, MAE, RRMSE, R^2 ; RMSE/MAE in mm dekad⁻¹). RRMSE uses each site's pixel-weighted mean L3 ET. In-domain hold-outs: Baixo 2025, Lamego 2022, JEN 2025 (a training site's unseen future year); transfer sites (MIT/KOGA/KMW) are fully held out. Best per (site, metric) in **bold**.

Table A1. Full per-site accuracy (RMSE, MAE, RRMSE, R^2) for all five models; bold = best per (site, metric).

Site (mean ET)	Model	RMSE	MAE	RRMSE %	R^2
Baixo (19.8)	Ensemble	3.82	2.66	19.3	0.879
	SwinIR	3.91	2.75	19.8	0.873
	Prithvi	3.99	2.75	20.2	0.867
	PRSN	4.93	3.13	24.9	0.789
	PRSN-PM	5.32	3.42	26.9	0.765
Lamego (22.5)	Ensemble	4.11	2.94	18.3	0.869
	SwinIR	4.10	2.92	18.2	0.869
	Prithvi	4.24	3.06	18.9	0.860
	PRSN	4.33	3.10	19.3	0.855
	PRSN-PM	4.40	3.10	19.6	0.850
Tunisia (JEN) (13.5)†	Ensemble	3.01	2.12	22.4	0.909
	SwinIR	3.00	2.10	22.3	0.910
	Prithvi	3.23	2.28	24.0	0.896

Site (mean ET)	Model	RMSE	MAE	RRMSE %	R ²
MIT (14.7)	PRSN	3.58	2.49	26.6	0.872
	PRSN-PM	3.58	2.54	26.6	0.872
	Ensemble	3.47	2.51	23.7	0.839
	SwinIR	3.56	2.58	24.3	0.831
	Prithvi	3.66	2.63	25.0	0.821
KOGA (24.5)	PRSN	5.09	3.70	34.7	0.703
	PRSN-PM	4.50	3.23	30.7	0.755
	Ensemble	4.06	3.12	16.6	0.844
	SwinIR	4.35	3.35	17.8	0.821
	Prithvi	4.21	3.23	17.2	0.832
KMW (25.1)	PRSN	5.36	3.54	21.8	0.754
	PRSN-PM	5.34	3.67	21.8	0.756
	Ensemble	4.33	3.30	17.3	0.800
	SwinIR	4.57	3.49	18.2	0.779
	Prithvi	4.37	3.33	17.4	0.798
	PRSN	5.64	4.43	22.5	0.638
	PRSN-PM	5.05	3.91	20.1	0.735

Across all six sites a direct model is best on every metric (Ensemble at five sites; SwinIR at Lamego and JEN). The PM anchor narrows the physics gap on the transfer sites — most visibly at MIT (RRMSE 34.7 %→30.7 %) and KMW (22.5 %→20.1 %) — but never overtakes the direct models.

Appendix B — Model configuration, hardware and runtime

Hardware and software. All models were trained and evaluated on a single **NVIDIA GeForce RTX 5090 (32 GB)** workstation (driver 596.36, CUDA 13.0) running Windows 11, with **PyTorch 2.12** (Python 3.12, conda environment). Training used mixed-precision (AMP), the AdamW optimiser and a OneCycle learning-rate schedule; data acquisition and compositing used Google Earth Engine and Python (§2.3). Times below are wall-clock on this single GPU. Table B1 lists each model’s configuration, trainable-parameter count and training cost.

Table B1. *Model configurations, trainable parameters and training cost (7-site training run, 100 epochs unless noted). The ensemble is a post-hoc 50/50 average and is not trained.*

Model	Architecture / key configuration	Trainable params	Epochs	Train time (RTX 5090)
SwinIR (direct)	Swin-transformer SR; embed dim 96, depths 4·4·4·4,	≈ 1.8 M	100	≈ 7.5 h (270 s/epoch)

Model	Architecture / key configuration	Trainable params	Epochs	Train time (RTX 5090)
Prithvi-V1 (direct)	heads 6·6·6·6, window 16, patch 256, 9-ch input IBM/NASA Prithvi geospatial ViT (~300 M); backbone frozen to layer 22, regression head trained (lr-head 5e-5, lr-backbone 1e-6)	head only (~few M of ~300 M)	100	≈ 2.9 h (105 s/epoch)
Ensemble	50/50 per-pixel mean of SwinIR + Prithvi (patch 256, overlap 64)	—	—	inference only
PRSN PRSN+Budyko	/ SwinIR-e96 texture backbone + δ -head + block-softmax W (block 16) + energy ceiling + advection gate (+Budyko regime attention)	≈ 1.8 M	100	≈ 1.6 h (58 s/epoch)
PRSN-PM	as PRSN, energy term = FAO-56 Penman–Monteith PET soft-hinge anchor (ERA5 + SRTM)	≈ 1.8 M	100	≈ 1.6 h (58 s/epoch)
Per-site fine-tune	SwinIR-e96 initialised from the 7-site model; lr 5e-5, batch 2, patch 256	≈ 1.8 M	30	≈ 0.5–1.6 h per site

All training runs used the same nine 20 m input channels and the seven-site training set (Study areas; Model inputs). Inference (full-scene Hann-blend tiling, patch 256 / overlap 64) runs in seconds-to-minutes per dekad per site on the same GPU.

Code and data availability

All input datasets are public: WaPOR L1/L3 ET (FAO WaPOR v3 object store), Sentinel-2 (Copernicus/Google Earth Engine), ERA5-Land, CHIRPS, SRTM, and ESA WorldCover (via Google Earth Engine), as cited in §2.3 (Table 2). The model-comparison code, comprising the data-preparation pipeline, the SwinIR/Prithvi/PRSN training and evaluation scripts, the fair per-dekad and mass-conservation evaluators, and the figure-generation scripts, is available at the project repository (https://github.com/iwmihq/wapor_africa). An interactive demonstrator of the downscaler, which generates 20 m ETa GeoTIFFs for a user-defined area of interest and WaPOR date, is publicly available as a Hugging Face Space (https://huggingface.co/spaces/IWMIHQ/WaPOR_20_m_Downscaler). Trained model checkpoints and the per-dekad evaluation artefacts underlying Tables 6–10 and Figures 4–6 are available from the authors on reasonable request.

Declarations

Funding. This work was funded by the Government of the Netherlands through the Food and Agriculture Organization of the United Nations (FAO), under FAO Project Symbol GCP/INT/729/NET and Budget Code TF.NSLDD.TFNL110020514. The associated administrative records include GRMS Supplier Number 38329 and PO Number 351633.

Author contributions. Z.K.: conceptualisation, methodology, software, analysis, writing – original draft. P.S.: supervision, writing – review and editing. M.G.A.: supervision, writing – review and editing. K.B.: writing – review and editing.

Competing interests. The authors declare no competing interests.

References

- Allen RG, Pereira LS, Raes D, Smith M (1998) Crop evapotranspiration — Guidelines for computing crop water requirements. FAO Irrigation and Drainage Paper 56. FAO, Rome.
- Anderson MC, Allen RG, Morse A, Kustas WP (2012) Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources. *Remote Sensing of Environment* 122:50–65. <https://doi.org/10.1016/j.rse.2011.08.025>
- Bastiaanssen WGM, Menenti M, Feddes RA, Holtslag AAM (1998) A remote sensing surface energy balance algorithm for land (SEBAL): 1. Formulation. *Journal of Hydrology* 212–213:198–212. [https://doi.org/10.1016/S0022-1694\(98\)00253-4](https://doi.org/10.1016/S0022-1694(98)00253-4)
- Beucler T, Pritchard M, Rasp S, Ott J, Baldi P, Gentine P (2021) Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters* 126(9):098302. <https://doi.org/10.1103/PhysRevLett.126.098302>

- Blatchford ML, Mannaerts CM, Njuki SM, Nouri H, Zeng Y, Pelgrum H, Wonink S, Karimi P (2020) Evaluation of WaPOR V2 evapotranspiration products across Africa. *Hydrological Processes* 34(15):3200–3221. <https://doi.org/10.1002/hyp.13791>
- Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. (2021) On the opportunities and risks of foundation models. arXiv:2108.07258. <https://doi.org/10.48550/arXiv.2108.07258>
- Carlson TN, Ripley DA (1997) On the relation between NDVI, fractional vegetation cover, and leaf area index. *Remote Sensing of Environment* 62(3):241–252. [https://doi.org/10.1016/S0034-4257\(97\)00104-1](https://doi.org/10.1016/S0034-4257(97)00104-1)
- Che X, Zhang HK, Sun Q, Ouyang Z, Liu J (2022) MODIS evapotranspiration downscaling using a deep neural network trained using Landsat 8 reflectance and temperature data. *Remote Sensing* 14(22):5876. <https://doi.org/10.3390/rs14225876>
- Claverie M, Ju J, Masek JG, Dungan JL, Vermote EF, Roger J-C, Skakun SV, Justice C (2018) The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment* 219:145–161. <https://doi.org/10.1016/j.rse.2018.09.002>
- Dong C, Loy CC, He K, Tang X (2015) Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(2):295–307. <https://doi.org/10.1109/TPAMI.2015.2439281>
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. (2021) An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*. arXiv:2010.11929. <https://doi.org/10.48550/arXiv.2010.11929>
- Drusch M, Del Bello U, Carlier S, Colin O, Fernandez V, Gascon F, et al. (2012) Sentinel-2: ESA’s optical high-resolution mission for GMES operational services. *Remote Sensing of Environment* 120:25–36. <https://doi.org/10.1016/j.rse.2011.11.026>
- FAO (2017) The future of food and agriculture — Trends and challenges. Food and Agriculture Organization of the United Nations, Rome.
- FAO (2018) WaPOR Database Methodology: Level 1, 2 and 3. Food and Agriculture Organization of the United Nations, Rome.
- Farr TG, Rosen PA, Caro E, Crippen R, Duren R, Hensley S, et al. (2007) The Shuttle Radar Topography Mission. *Reviews of Geophysics* 45(2):RG2004. <https://doi.org/10.1029/2005RG000183>
- Funk C, Peterson P, Landsfeld M, Pedreros D, Verdin J, Shukla S, et al. (2015) The climate hazards infrared precipitation with stations — a new environmental record for monitoring extremes. *Scientific Data* 2:150066. <https://doi.org/10.1038/sdata.2015.66>
- Gao B-C (1996) NDWI — A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment* 58(3):257–266. [https://doi.org/10.1016/S0034-4257\(96\)00067-3](https://doi.org/10.1016/S0034-4257(96)00067-3)

- He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2022) Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 16000–16009. <https://doi.org/10.1109/CVPR52688.2022.01553>
- Jakubik J, Roy S, Phillips CE, Fraccaro P, Godwin D, Zadrozny B, et al. (2023) Foundation models for generalist geospatial artificial intelligence (Prithvi). *arXiv:2310.18660*. <https://doi.org/10.48550/arXiv.2310.18660>
- Karimi P, Bastiaanssen WGM, Molden D (2013) Water Accounting Plus (WA+) — A water accounting procedure for complex river basins based on satellite measurements. *Hydrology and Earth System Sciences* 17:2459–2472. <https://doi.org/10.5194/hess-17-2459-2013>
- Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, Ganguly A, Shekhar S, Samatova N, Kumar V (2017) Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering* 29(10):2318–2331. <https://doi.org/10.1109/TKDE.2017.2720168>
- Ke Y, Im J, Park S, Gong H (2016) Downscaling of MODIS one kilometer evapotranspiration using Landsat-8 data and machine learning approaches. *Remote Sensing* 8(3):215. <https://doi.org/10.3390/rs8030215>
- Kiala Z, Schmitter P, Garcia Andarcia M (2026) Downscaling dekadal WaPOR actual evapotranspiration from 300 m to 20 m for irrigation monitoring. *Research Square preprint (under review)*. <https://doi.org/10.21203/rs.3.rs-9687804/v1>
- Liang J, Cao J, Sun G, Zhang K, Van Gool L, Timofte R (2021) SwinIR: Image restoration using Swin Transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp 1833–1844. <https://doi.org/10.1109/ICCVW54120.2021.00210>
- Lim B, Son S, Kim H, Nah S, Lee KM (2017) Enhanced deep residual networks for single image super-resolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp 136–144. <https://doi.org/10.1109/CVPRW.2017.151>
- Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin Transformer: Hierarchical Vision Transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp 10012–10022. <https://doi.org/10.1109/ICCV48922.2021.00986>
- Molden D, Oweis T, Steduto P, Kijne J, Hanjra M, Bindraban P (2010) Improving agricultural water productivity: Between optimism and caution. *Agricultural Water Management* 97(4):528–535. <https://doi.org/10.1016/j.agwat.2009.03.023>
- Muñoz-Sabater J, Dutra E, Agustí-Panareda A, Albergel C, Arduini G, Balsamo G, et al. (2021) ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data* 13:4349–4383. <https://doi.org/10.5194/essd-13-4349-2021>
- Peng J, Loew A, Merlin O, Verhoest NEC (2017) A review of spatial downscaling of satellite remotely sensed soil moisture. *Reviews of Geophysics* 55(2):341–366. <https://doi.org/10.1002/2016RG000543>

- Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat (2019) Deep learning and process understanding for data-driven Earth system science. *Nature* 566:195–204. <https://doi.org/10.1038/s41586-019-0912-1>
- Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS 9351, pp 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
- Senay GB, Bohms S, Singh RK, Gowda PH, Velpuri NM, Alemu H, Verdin JP (2013) Operational evapotranspiration mapping using remote sensing and weather datasets: A new parameterization for the SSEB approach. *Journal of the American Water Resources Association* 49(3):577–591. <https://doi.org/10.1111/jawr.12057>
- Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. *Journal of Big Data* 6:60. <https://doi.org/10.1186/s40537-019-0197-0>
- Singh RK, Senay GB, Velpuri NM, Bohms S, Verdin JP (2014) On the downscaling of actual evapotranspiration maps based on combination of MODIS and Landsat-based actual evapotranspiration estimates. *Remote Sensing* 6(11):10483–10509. <https://doi.org/10.3390/rs6110483>
- Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Qiao Y, Loy CC (2018) ESRGAN: Enhanced super-resolution generative adversarial networks. *European Conference on Computer Vision Workshops (ECCVW)*, LNCS 11133, pp 63–79. https://doi.org/10.1007/978-3-030-11021-5_5
- Willard J, Jia X, Xu S, Steinbach M, Kumar V (2022) Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys* 55(4):1–37. <https://doi.org/10.1145/3514228>
- WWAP (2021) The United Nations World Water Development Report 2021: Valuing Water. UNESCO World Water Assessment Programme, Paris.
- Xie H, You L, Wielgosz B, Ringler C (2014) Estimating the potential for expanding smallholder irrigation in Sub-Saharan Africa. *Agricultural Water Management* 131:183–193. <https://doi.org/10.1016/j.agwat.2013.08.011>
- Yang Y, Long D, Guan H, et al. (2020) Investigating impacts of drought and disturbance on evapotranspiration over global forests in recent two decades. *Remote Sensing of Environment* 248:111415. <https://doi.org/10.1016/j.rse.2018.12.017>
- You L, Ringler C, Wood-Sichra U, Robertson R, Wood S, Zhu T, Nelson G, Guo Z, Sun Y (2011) What is the irrigation potential for Africa? A combined biophysical and socioeconomic approach. *Food Policy* 36(6):770–782. <https://doi.org/10.1016/j.foodpol.2011.09.001>
- Zanaga D, Van De Kerchove R, Daems D, De Keersmaecker W, Brockmann C, Kirches G, et al. (2022) ESA WorldCover 10 m 2021 v200 [Data set]. European Space Agency. <https://doi.org/10.5281/zenodo.7254221>

Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image super-resolution using very deep residual channel attention networks. European Conference on Computer Vision (ECCV), LNCS 11211, pp 286–301. https://doi.org/10.1007/978-3-030-01234-2_18