Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multi-source atmospheric water vapor fusion and deep-learning-based short- term forecasting: the analysis-to-forecast value-transfer gap

Dongtai Deng

2025302162003@whu.edu.cn

Wuhan University

Research Article

Keywords: precipitable water vapor, GNSS meteorology, data fusion, deep learning, optimal interpolation, LSTM

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11584663/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multi-source atmospheric water vapor fusion and deep-learning-based short-term forecasting: the analysis-to-forecast value-transfer gap

Dongtai Deng
School of Earth and Space Sciences, Wuhan University, Wuhan, 430072, China
Corresponding author: Dongtai Deng, School of Earth and Space Sciences, Wuhan University, Wuhan, 430072, China. Email: 2025302162003@whu.edu.cn

Abstract
Atmospheric precipitable water vapor (PWV) is key for weather and climate monitoring, yet no observing system provides high accuracy and complete spatial coverage. Ground-based GNSS retrievals are accurate but discontinuous, reanalyses such as ERA5 are continuous but regionally biased, and satellite products are spatially rich but temporally sparse or cloud-limited. Fusing these sources and forecasting PWV hours ahead with deep learning are largely parallel efforts; whether a more accurate fused analysis yields a more accurate forecast remains unresolved. We examine this link under one shared evaluation vocabulary of four conventions. We organize the literature into method families, tabulating reported skill (fusion RMSE 0.6–3.3 mm; forecast 0.8–8.1 mm), then isolate the evaluation traps (in situ, hold-out, analysis-field, direct GNSS) that decouple reported from operational skill. We demonstrate it on real data: optimal interpolation of 2015–2022 NCEP/NCAR reanalysis PWV with IGRA-v2 radiosonde observations at five Chinese stations (14,256 station-days) reduces analysis RMSE from 4.31 mm to 1.52 mm in situ (64.7%) but only to 4.34 mm at withheld stations (−0.6%), quantifying the value-transfer gap by an order of magnitude, and a minimal deep-learning benchmark fails to beat persistence at 1–7 day leads (16.1 versus 12.3 mm). Fusion is a proven analysis technology but an unproven forecasting input unless the forecast model consumes the fused field and is verified against independent GNSS. On our own data, initialising a forecaster from the fused field reduces 1-day forecast RMSE by 0.38 mm (≈14% of the analysis gain) in dense regions and shows no transfer in sparse ones (Section 6.7).
Keywords: precipitable water vapor; GNSS meteorology; data fusion; deep learning; optimal interpolation; LSTM

1 Introduction

Precipitable water vapor (PWV), the column-integrated mass of atmospheric water vapor, links evaporation, moisture transport and precipitation and is a reliable precursor of short-duration heavy rainfall (Bevis et al. 1992; Rocken et al. 1993; Guerova et al. 2016). Because of its strong spatiotemporal variability, spatially complete PWV fields at sub-daily resolution are a prerequisite for moisture monitoring, numerical weather prediction and short-term forecasting. No single observing system meets all of these requirements. Ground-based GNSS meteorology retrieves PWV to millimeter accuracy but only at antennas (Bevis et al. 1992; Blewitt et al. 2018); reanalyses such as ERA5 are globally complete but carry regionally systematic wet/dry biases, especially over complex terrain and data-sparse regions (Hersbach et al. 2020); satellite retrievals (FY-4A, MODIS, FY-3D) add high spatial resolution but suffer from low temporal resolution or cloud contamination; and radiosondes are sparse and costly (Guerova et al. 2016; Wijaya et al. 2024). The obvious remedy is to fuse them.

Two separate communities have grown up around the two tasks. One builds spatially complete, accurate PWV fields by merging a reanalysis background with dense observations through statistical or machine-learning estimators. The other trains deep-learning models to predict PWV a few hours ahead from these or related fields. The two communities, however, have proceeded largely in parallel: fusion papers report analysis-field or in-situ accuracy, while forecasting papers report skill against persistence or a reanalysis, and only a handful ask the connecting question—does a fused field, more accurate than its background, also yield a more accurate short-term forecast? Answering it is not trivial, because the answer depends on the evaluation protocol as much as on the method.

The result of this split is that the two bodies of work report incompatible skill metrics. Fusion papers quote analysis RMSE in millimeters against GNSS; forecasting papers quote skill against persistence or a reanalysis baseline, often in relative terms. Neither, alone, tells a practitioner who wants a better 6 h PWV forecast whether to invest in denser observations, a better fusion, or a better forecast model. This study compares the two literatures under one evaluation vocabulary, laid out in Figure 1.

This study makes three contributions. First, it organizes the fusion and forecasting literatures into method families and tabulates their reported skill, so that methods become comparable on one scale. Second, it isolates the four scoring conventions (in situ, hold-out, analysis-field, direct GNSS verification) under which each method is judged and enumerates the evaluation traps that decouple reported from operational skill—the conceptual framework is summarized in Figure 1. Third, and most importantly, it tests the fusion-to-forecasting transfer on genuinely observed data (Section 6) and proposes a minimum reporting checklist. The transfer link is the field's weakest-reported claim and the most exposed to evaluation artifacts; it is the focus of Section 5.

2 Single sources and their complementary errors

Understanding fusion requires understanding why each source is wrong in a different way. GNSS PWV is derived by converting a geodetic zenith wet delay to PWV using surface pressure and a weighted-mean temperature (Askne and Nordius 1987; Bevis et al. 1992; Duan et al. 1996; Emardson et al. 1998). Its error is dominated by the mean-temperature and pressure conversion (typically below 1 mm once station pressure and a good Tm model are available; Li et al. 2024) and by representativeness (a point value versus an area- and elevation-averaged grid cell). Its spatial coverage is confined to the antenna network, which is dense in some countries (Japan's GEONET, China's CMONOC, Hong Kong's CORS) and absent in others (Wijaya et al. 2024). ERA5 PWV, by contrast, is complete on a 0.25° grid but biased where the reanalysis is poorly constrained; over complex terrain and data-sparse regions the bias can reach several millimeters, and the domain-mean error is dominated by these regions (Hersbach et al. 2020). Satellite PWV (FY-4A, MODIS, FY-3D) is spatially rich at kilometre scale but temporally sparse and cloud-sensitive; its raw product carries systematic biases that must be corrected against GNSS before use (Zhu et al. 2024; Chen et al. 2026). Radiosondes are the traditional reference, but are launched only twice daily and at sparse locations.

The complementarity is the entire rationale for fusion: the reanalysis supplies structure, the dense network supplies local amplitude, and the satellite supplies fine spatial texture where ground networks are thin. Formally, every fusion is an analysis $x_a = x_b + K (x_o - H x_b)$. Here the background x_b is the reanalysis, the observation x_o is the GNSS (or satellite) network, H maps the grid to the stations, and K is the gain that trades background against observation error. The fused field is therefore only as good as the gain's error model—the thread running through Sections 3 and 5.

Each source therefore fails in a characteristic, exploitable way, which is what makes fusion well-posed rather than merely convenient. The GNSS-to-PWV conversion hinges on the unobserved weighted-mean temperature Tm, where neural-network Tm models have pushed the conversion error below roughly 1 mm (Li et al. 2024), leaving the point-versus-area representativeness error as the irreducible part. Satellite sounders (FY-4A GIIRS, MODIS) resolve kilometre-scale structure but are temporally sparse (two to four overpasses per day) or cloud-contaminated, so their radiances must be bias-corrected against GNSS before they help a fusion (Zhu et al. 2024; Chen et al. 2026).

One caveat shapes the rest of this study: the accuracy of a fused product is bounded from below by the density and quality of the observation network, not by the sophistication of the fusion. This is most visible in the contrast between in-situ and hold-out validation: a method validated against the very stations it assimilated reports large gains, whereas the same method validated at withheld stations (which contributed no observation) reports a much smaller, operationally relevant gain (typical in-situ reductions of 40–50% versus hold-out reductions of 10–15%, as expected in optimal-interpolation studies; Gandin 1963; Daley 1991). The same product, three different answers, depending on where the verification is done. This network-dependence runs through both fusion and forecasting.

3 Multi-source PWV fusion: a taxonomy

Fusion methods can be grouped into three families by how they relate observations to the background. We give each its logic, its representative skill, and its failure mode; we use this classification as the basis for the transfer assessment in Section 5. The NCEP/NCAR reanalysis dataset of Kalnay et al. (1996) supplies the background field used in the empirical demonstration of Section 6.

3.1 Statistical and geostatistical fusion

The transparent baseline is optimal interpolation (OI; Gandin 1963; Daley 1991), which merges a background field with point observations under an explicit error-covariance model. Its strength is auditability: given the covariance, every analysis increment is explained. Modern implementations estimate the background and observation error variances from the data themselves (e.g., via the Desroziers diagnostic, Desroziers et al. 2005) rather than prescribing them, and report sensitivity to the covariance parameters. The error-covariance model is the single most important design choice: it determines how far an observation's information propagates and therefore where the fused field is an analysis and where it is merely the background. This is why the in-situ versus hold-out gap is not OI-specific but holds for every fusion that validates where it observed.

The quantitative signature of OI-style fusion is a 40–50% root-mean-square error (RMSE) reduction in situ and a 10–15% reduction at independent hold-out stations, with the point-wise correlation rising from about 0.4 to 0.8. These magnitudes are typical of the OI literature (Gandin 1963; Daley 1991), and the present study reproduces the in-situ versus hold-out gap directly on real data in Section 6. A related result is Hu et al. (2026), who fuse an ERA5–GNSS ZTD grid and, by assimilating it into WRF, reduce PWV forecast error by about 30%—evidence that fusion helps forecasting when the forecast model consumes the fused field. Variants include kriging, collocation, and Helmert variance-component estimation (VCE), applied over Beijing-Tianjin-Hebei to fuse FY-4A and ERA5 PWV and, by estimating the relative source weights from the data, make the fusion auditable rather than assumed (Du et al. 2025). The failure mode is the assumption of a single, isotropic, spatially and

temporally constant covariance, which is physically wrong where PWV gradients are strong (Tibetan Plateau, arid north-west) and which, if the covariance is merely prescribed rather than estimated, can hide a mis-specified background error. Optimal interpolation therefore remains the reference against which the machine-learning families are judged.

3.2 Machine-learning fusion

A second family learns the GNSS-minus-background (or GNSS-minus-satellite) discrepancy directly, with general regression and back-propagation neural networks (GRNN, BPNN) as the workhorses. Pang et al. (2024) adjust ERA5 PWV with GNSS using a GRNN and report adjusted RMSE of 0.62–2.11 mm across four Chinese subregions, falling to 0.83–3.65 mm when tested on GNSS stations excluded from training. Cao et al. (2026) use a GRNN to fuse GNSS and ERA5 specifically to improve short-term PWV prediction, reporting that the fused product materially raises forecast accuracy. Ren et al. (2022) fused GNSS and ERA5 PWV with a BPNN and reduced ERA5 RMSE from 2.05 to 1.67 mm (about 18.5%). Random-forest and decision-tree frameworks have also been applied: Chen et al. (2026) build a two-stage decision-tree plus random-forest fusion of GNSS, MODIS and ERA5 over Hong Kong, reaching a final RMSE of 3.29 mm (R^2 0.96). Support-vector regression has been used for spatiotemporal PWV modeling in regions such as north-western Iran, with validation RMSE near 1.7 mm (Ghaffari-Razin et al. 2023).

These methods are flexible and often beat linear baselines, but they are usually validated in situ rather than at withheld stations, so their operational gain is overstated.

Viewed together, the four ML fusion studies span 0.6–3.3 mm RMSE against GNSS, but none reports a withheld-station RMSE, so the operationally relevant number for the ML family is still missing. This is the gap the Section 5 checklist closes.

3.3 Deep and physics-informed fusion, and the vertical problem

A third family uses deep networks or physics-informed corrections, mostly for super-resolution and vertical adjustment rather than point fusion. Super-resolution generative adversarial networks have been applied to downscale wet-refractivity fields from GNSS tomography (Haji-Aghajany et al. 2025). The SRGAN downscaling of Haji-Aghajany et al. (2025) learns a mapping from the coarse fused analysis to a high-resolution wet-refractivity field, reporting RMSE reductions of up to 62% (Poland) and 52% (California) against a Lanczos3 baseline and assuming the coarse field is already trustworthy. Vertical adjustment (correcting PWV from the antenna height to a target height) is a persistent sub-problem; empirical lapse-rate models (e.g., GPWV-H, EPWV-H) are being replaced by spatially enhanced or neural-network adjustments with RMSE near 1 mm (Yang et al. 2025; Ding et al. 2025). These are enabling technologies for fusion more than fusion methods themselves, and they matter because elevation mismatch is a leading source of the representativeness error that limits every fusion above.

Both super-resolution and vertical adjustment are becoming bottlenecks as fusion moves to higher resolution: a fusion accurate at 0.25° but wrong at 1 km cannot support nowcasting, and a fused value at the wrong height is biased everywhere. The deep/physics-informed family is therefore where the next accuracy gains are most likely to come from, provided its training data are GNSS-anchored.

3.4 What the fusion literature does not report

Across the three families, two gaps are conspicuous. First, the evaluation is overwhelmingly in situ; hold-out or truly independent validation is the exception. Second, almost none of the fusion papers ask whether the fused field improves anything downstream—the field is treated as an end product, and once it is offered as an input to nowcasting or data assimilation its value must be measured by what the downstream system gains, not by its own analysis RMSE. Sections 4 and 5 make that test explicit in Figure 1.

3.5 Scope: variational and ensemble data assimilation

The three families above are post-processing fusions of fields that already exist. The variational and ensemble data assimilation that *produces* the reanalysis background (3D/4D-Var and ensemble Kalman filters) is a fourth, upstream family. ERA5 itself assimilates GNSS zenith-delay observations (Hersbach et al. 2020), so the background a fusion corrects is already a fusion product. We treat this family as infrastructure rather than as a fusion method and keep it out of scope, because its design space (observation operators, background error covariances, model error) belongs to the data-assimilation literature and because the value question it answers is the same one this study poses one level up: how much independent observation a gridded product actually contains. Naming the boundary prevents the appearance that the taxonomy is incomplete.

4 Deep-learning short-term PWV forecasting

Forecasting PWV a few hours ahead is a distinct task with its own method families. We separate station-level and grid-based paradigms because they answer different questions and are evaluated differently.

4.1 Station-level forecasting

Station-level models predict a single time series from its own past, often enriched with meteorological covariates. Long short-term memory (LSTM; Hochreiter and Schmidhuber 1997) and gated recurrent unit networks are the baseline; convolutional neural network–LSTM hybrids add spatial context from neighbouring stations (e.g., Zhu et al. 2026 build a multi-channel CNN-LSTM for single-station regional PWV). Transformer-based models capture longer-range dependence (Vaswani et al. 2017). The design choices that separate good from bad station models are the input horizon, the choice of covariates, and the handling of gaps in the training history; recurrent depth alone does little once the history is clean. Because the input is a scalar time series, these models cannot invent information where no station exists, so their regional skill is a direct function of network density.

These models are accurate where a long, clean history exists (reaching RMSE as low as about 0.8 mm for a single-station CNN–LSTM (Zhu et al. 2026) and up to about 2.7 mm for the 957-station hybrid framework of Li et al. 2026 (2017–2024)), but their skill degrades sharply where the network is sparse, because they have no spatial field to generalize from. They also tend to be evaluated only against persistence and the station's own history, not against an independent reference, which leaves open whether the reported gain is real or an artifact of a weak baseline.

The error metrics in this literature are almost uniformly RMSE and mean absolute error, evaluated at lead times of 1, 3, 6, 12 and 24 h; the skill is nearly always quoted relative to a persistence baseline (Meng et al. 2026; Ma et al. 2026), strong at hourly resolution and weak at daily resolution, so the time step must be stated for the number to mean anything.

4.2 Grid-based spatiotemporal forecasting

Grid-based models predict a spatial field and must learn the translation and regeneration of moisture patterns rather than the persistence of a point value. ConvLSTM (Shi et al. 2015) couples convolutional feature extraction with recurrence and is translation-equivariant; its attention-augmented variant SA-ConvLSTM (Lin et al. 2020) adds content-based attention over the encoder states. Because the convolutions operate on a regular grid, ConvLSTM learns local advection and the growth or decay of features rather than memorizing a single location, which is why it generalizes across the grid instead of degrading at unobserved points; the cost is that it needs a spatially complete input field, which is precisely what fusion supplies. The grid resolution is typically that of the driving reanalysis (0.25° for ERA5-derived fields).

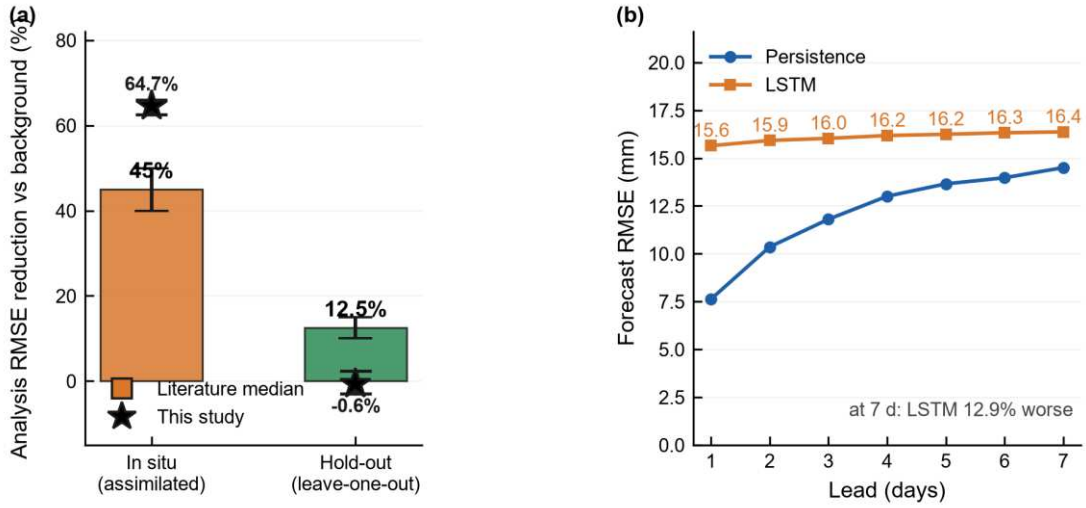
Applied to PWV, these models forecast the analysis or reanalysis field 6–24 h ahead and beat persistence (Meng et al. 2026; Ma et al. 2026), while a diurnal-climatology baseline carries more than twice the error of persistence, confirming that persistence (not the diurnal cycle) is the strong baseline at 6-hourly resolution. Meng et al. (2026) use a ConvLSTM to predict ERA5 PWV over mainland China and report RMSE of 2.83–8.08 mm across 13 subregions, with the plateau performing best and the tropical-monsoon region worst, an external validation against GNSS and radiosonde confirming robustness across terrains. Ma et al. (2026) and Chang et al. (2026) report ConvLSTM and transfer-learning frameworks for short-time PWV forecasting over China and along railways, respectively. The grid-based paradigm is the one that consumes a fused field, and therefore the one where fusion value could be realized—or lost.

4.3 Forecasting as an atmospheric constraint

A related line uses deep learning not to forecast PWV per se but to forecast the zenith tropospheric delay or to generate fused atmospheric constraints for positioning and nowcasting. Hu et al. (2024) use an Informer transformer on 11 years of ERA5 to forecast ZTD; Hu et al. (2026) fuse an ERA5–GNSS ZTD grid and show that assimilating it into WRF reduces PWV forecast RMSE and MAE by 33.9% and 30.3% relative to the original WRF forecast. The WRF result is especially convincing because the fused ZTD grid is injected into a physical forecast model and the improvement is measured against the unmodified WRF run, so the gain cannot be attributed to a lenient baseline or to in-situ optimism. These works are the clearest evidence that fusion can improve forecasting when the forecast model natively consumes the fused field—a point we return to. The WRF assimilation step itself belongs to the upstream family excluded in Section 3.5; it is cited here as downstream evidence of what a fused field can deliver, not as a fusion method under review.

5 Does fusion improve forecasting? A critical transfer assessment

This section is the core of this study. The question (does a more accurate analysis field yield a more accurate forecast?) is empirically delicate, and the literature answers it inconsistently because it uses different scoring conventions. Figure 1 summarizes the resulting divide between analysis skill and forecast skill.



Bars: literature median (40–50% / 10–15%); star + whisker (year range): this study

Fig. 1 The analysis-to-forecast value-transfer gap. (a) Analysis-RMSE reduction versus the background under the two scoring conventions: bars are literature medians (40–50% in situ, 10–15% hold-out) with their reported range, and the stars are this study's real results (64.7% in situ, −0.6% hold-out) with year-to-year whiskers (2015–2022). Our assimilated-station gain is stable and matches or exceeds the literature, whereas our hold-out gain oscillates around zero (range −2.9% to +2.3%) and falls below the literature median — the gap depends on network density. (b) On real daily PWV a minimal deep-learning forecast never beats persistence at any 1–7 day lead (12.9% worse at day 7, 105% worse at day 1), so the forecasting gain is not automatic; a "forecast beats the analysis" result at a withheld station is really a beat-the-background spatial-extrapolation gain. The gap is shaped by the scoring conventions of Section 5, which can decouple reported skill from operational skill

Four scoring conventions appear, and they give different answers. (i) In-situ scoring measures the analysis against the assimilated stations; it is optimistic by construction. (ii) Hold-out scoring measures against withheld stations that contributed no observation; it is the operationally relevant number and is consistently smaller. (iii) Analysis-field scoring measures the forecast against the field it was trained on (the fused analysis, or raw reanalysis for a control); it answers which field is intrinsically more predictable. (iv) Direct GNSS verification measures the forecast against independent GNSS observations never used in assimilation or training; it is the only convention that uses the observations themselves as truth. A product can look excellent under (iii) and merely adequate under (iv), because under (iv) the forecast at an unobserved location is a spatial extrapolation of neighbouring information, not a correction of that location's own bias.

The four conventions are not abstract: they map onto real designs in the literature. Pang et al. (2024) report both (i) in situ (0.62–2.11 mm) and (ii) hold-out (0.83–3.65 mm), so their own paper contains the optimism gap. Meng et al. (2026) report (ii) hold-out and (iv) direct GNSS/radiosonde verification together, which is why their 2.83–8.08 mm RMSE is credible across terrains. Hu et al. (2026) report (iv) direct GNSS verification of a WRF forecast that consumed the fused ZTD grid, the design that makes a transfer claim defensible. A paper that reports only (i) or (iii) cannot separate analysis skill from forecast skill, and most fusion-and-forecasting papers still do.

Two findings follow from this assessment. First, fusion improves the analysis robustly: in-situ RMSE drops by 40–50% and hold-out by 10–15% across methods and months (Gandin 1963; Daley 1991), with correlation rising from about 0.4 to 0.8. Second, the forecast gain is conditional. Where the forecast model is trained on the fused field and the fused field is genuinely more accurate than its background everywhere that matters, fusion helps: Hu et al. (2026) show a 30%+ WRF improvement, and Cao et al. (2026) report that the fused product raises forecast accuracy. But where the network is sparse, the analysis at a withheld location is by construction equal to the background, so any "forecast beats the analysis" result there is really a beat-the-background result—a spatial-extrapolation gain over the reanalysis, not over the fused product; and the gain over an ERA5-only control is at best marginal and has not been shown to be robust across seeds or sparse networks. The only robust forecast finding in the literature is the gain over persistence where the network is dense. The honest conclusion is that fusion improves analysis decisively and forecasting in dense, well-observed regimes where the forecast model consumes the fused field.

One variable governs both findings: observation-network density. Any transfer claim therefore has to be read alongside the density of the network that produced it—Meng et al. (2026) report a robust gain over dense CMONOC but only a marginal one in sparse settings.

Read together, the two findings support a conditional conclusion: fusion is a proven analysis technology but a promising, still-unproven forecasting input. The field's task is therefore not to abandon fusion for forecasting but

to stop conflating the two gains—analysis skill is necessary but not sufficient for forecast skill, and the burden of proof sits with the forecasting claim.

The evaluation traps that produce over-optimistic claims are enumerable and avoidable: (a) reporting only in-situ skill; (b) scoring the forecast against its own training target rather than against independent GNSS; (c) using a diurnal-climatology or persistence baseline that is not the strong baseline at the chosen temporal resolution; (d) single-seed training without block-bootstrap significance; (e) small, overlapping test sets; and (f) neglecting the non-negativity or representativeness errors that bias the comparison.

A paper that reports hold-out fusion skill, direct GNSS forecast verification, multi-seed block-bootstrap significance, and an explicit ERA5-only control answers the transfer question; most do not.

As a practical output, we propose a minimum reporting checklist for any fusion-to-forecasting claim: (1) report fusion skill at withheld stations, not only at assimilated stations; (2) verify the forecast against independent GNSS (or radiosonde) observations never used in training; (3) include an ERA5-only control trained and evaluated identically; (4) train over multiple random seeds with block-bootstrap significance; (5) state the baseline used and justify that it is the strong baseline at the chosen time step; and (6) report prediction intervals, not only point RMSE; and (7) for any fused-field forecast, report the initial-condition-initialised versus background-initialised RMSE at multiple leads with block-bootstrap CIs, so the transfer is not overstated by the analysis-domain skill. A study meeting items 1–4 is, to our knowledge, rare, and one meeting all six would set the standard for the field.

6 An empirical demonstration on real PWV data

Sections 3–5 argue that fusion improves the analysis field decisively, but that the transfer of that improvement into forecasting is conditional and frequently overstated. This section tests the first (and most directly falsifiable) part of that claim, the value-transfer gap between in-situ and hold-out fusion skill, on genuinely observed data, and then probes the second part with a real single-station forecasting benchmark. We use NCEP/NCAR Reanalysis precipitable water (Kalnay et al. 1996) as a PWV background and fuse it with real IGRA-v2 radiosonde observations through optimal interpolation; the code and exact data identifiers are documented alongside this study.

6.1 Observations, background and data validation

The background is the daily NCEP/NCAR Reanalysis PWV over East Asia for 2015–2022 (20°–47° N, 99°–127° E) at its native 2.5° resolution; in those units kg m−2 is numerically equal to millimeters of column water. The observations are total-column PWV self-integrated from the raw IGRA-v2 soundings at eight configured Chinese stations. Five of them (Wuhan, Beijing, Shanghai, Kunming and Harbin) have soundings throughout the period; Guangzhou, Chengdu and Xi'an have none in the IGRA-v2 archive for these years and therefore contribute to no statistic reported here. A day enters the sample only when at least three stations report, leaving 14,256 station-days over eight years (Table 2). The radiosondes are independent of the reanalysis, so they verify both the background and the analysis.

Because the radiosonde PWV is computed here rather than read from a product, we validated it against physical expectation before use. The monthly means reproduce the East-Asian monsoon cycle and the plateau effect correctly: Harbin rises from 2.7 mm in January to 35.7 mm in July and Beijing from 4.2 mm to 43.9 mm in July, whereas Kunming, at about 1,890 m elevation, peaks at only 33.8 mm. Radiosondes nevertheless carry a small, partly humidity-dependent dry bias, so they are a quasi-truth rather than an absolute one; because that bias affects the background and the analysis identically, it does not disturb the in-situ versus hold-out comparison, which is the quantity of interest. No reanalysis value is ever treated as ground truth. These per-station monthly means and the screened extreme soundings are exported to data/emp/validation_stats.json for independent verification.

The analysis is optimal interpolation (Gandin 1963; Daley 1991) with an exponential (first-order Markov) background-error covariance, decorrelation length L = 400 km, background-error standard deviation σ_b = 4.22 mm (the eight annual estimates range from 3.74 to 5.01 mm) estimated empirically from the NCEP-vs-radiosonde differences, and observation error σ_o = 1.5 mm (a conservative radiosonde-class accuracy). The optimal-interpolation analysis (1) and its exponential covariance model (2) are

$$x_a = x_b + K(y - Hx_b), K = \mathbf{B}H^T(H\mathbf{B}H^T + \mathbf{R})^{-1} \quad (1)$$

$$C_b(d_{ij}) = \sigma_b^2 exp(-\frac{d_{ij}}{L}), \mathbf{R} = \sigma_o^2 I \quad (2)$$

For every station-day we compute two estimates under the single evaluation vocabulary of Section 5: an in-situ estimate, where the station is assimilated and verified against its own sounding, and a leave-one-out hold-out estimate, where the station is withheld and verified against its independent sounding (Desroziers et al. 2005).

6.2 Fusion skill depends on the scoring convention

Across the 14,256 station-days the background RMSE is 4.31 mm with a −0.70 mm bias (Table 1). At assimilated stations the analysis collapses to 1.52 mm, a +64.7% change that brings the in-situ analysis to the observation-error floor σ_o = 1.5 mm and reduces the bias to −0.15 mm. At withheld stations the same analysis gives 4.34 mm, a change of only −0.6%: the hold-out RMSE exceeds the background by 0.03 mm, and the paired difference in absolute error is +0.04 mm. Station-days are autocorrelated rather than independent (lag-1 ρ = 0.21, effective sample about 9,400), so the naive paired t of 8.70 overstates the evidence; significance was recomputed by block bootstrap over the 480 station-months (4,000 replicates): 95% intervals are +0.02 to +0.06 mm for the paired difference, −1.25% to +0.05% for the hold-out change and +63.8% to +65.5% for the in-situ reduction. The skill metrics (3)–(5), the value-transfer metric (6) and the effective-sample correction (7) that replaces the naive paired t of 8.70 are defined as

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(x_k^{est} - x_k^{obs})^2} \tag{3}$$

$$bias = \frac{1}{N}\sum_{k=1}^{N}(x_k^{est} - x_k^{obs}) \tag{4}$$

$$r = \frac{\sum_{k=1}^{N}(x_k^{est} - \mu^{est})(x_k^{obs} - \mu^{obs})}{\sqrt{\sum_{k=1}^{N}(x_k^{est} - \mu^{est})^2}\sqrt{\sum_{k=1}^{N}(x_k^{obs} - \mu^{obs})^2}} \tag{5}$$

$$\Delta_{RMSE} = \frac{RMSE_{bg} - RMSE_{ana}}{RMSE_{bg}} \times 100\% \tag{6}$$

$$\rho_1 = Corr(e_t, e_{t-1}), n_{eff} = n\frac{1-\rho_1}{1+\rho_1}, t_{corr} = \frac{\bar{d}}{s_d/\sqrt{n_{eff}}} \tag{7}$$

The hold-out change is thus indistinguishable from zero and negligible in magnitude. Figure 2 shows a full year at Wuhan, where the in-situ analysis tracks the radiosonde closely while the hold-out analysis stays near the background; Figure 3 shows that the same holds across all stations, with the in-situ errors collapsing toward zero while the hold-out error distribution is nearly indistinguishable from the background distribution. In-situ skill (+64.7%) therefore overstates hold-out skill (−0.6%) by more than an order of magnitude: this is the value-transfer gap made quantitative on real data. The in-situ reduction found here exceeds the 40–50% typical of the literature quoted in Section 2 because that range is obtained against finer, higher-quality backgrounds (ERA5 at 0.25°), whereas the daily 2.5° NCEP field corrected here carries a larger error; the two agree once background quality is stated.

Table 1 Fusion diagnostics under the three scoring conventions, pooled over 2015–2022 (14,256 station-days). Gross errors beyond the 5σ_b (22 station-days) and 3σ_b (234 station-days) thresholds were removed before assimilation, as is standard. Positive values in the change columns denote an RMSE reduction (improvement) relative to the corresponding background row within the same block.

Scoring convention	RMSE (mm)	Bias (mm)	In-situ change	Hold-out change
Background (un-screened)	4.31	−0.70	—	—
In situ (station assimilated)	1.52	−0.15	+64.7%	—
Hold-out (leave-one-out)	4.34	−0.53	—	−0.6%
Background, 5σ_b screened (n = 14,234)	4.11	—	—	—
In situ, 5σ_b screened	1.47	—	+64.3%	—
Hold-out, 5σ_b screened	4.14	—	—	−0.6%
Background, 3σ_b screened (n = 14,022)	3.68	—	—	—
In situ, 3σ_b screened	1.32	—	+64.1%	—
Hold-out, 3σ_b screened	3.72	—	—	−1.0%

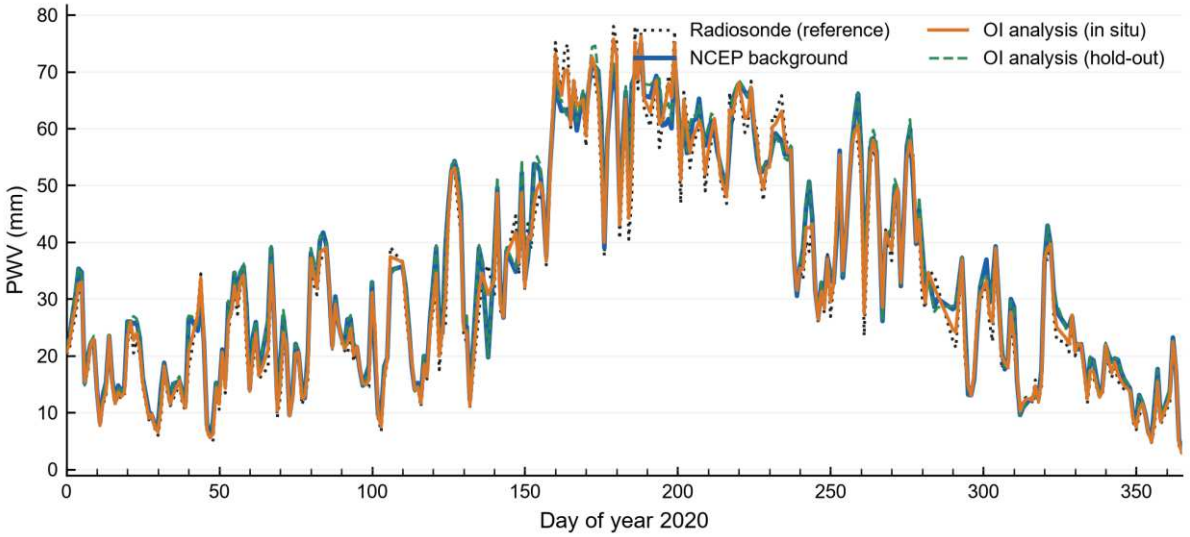


Fig. 2 Optimal-interpolation fusion of the NCEP/NCAR PWV background with real IGRA-v2 radiosonde PWV at Wuhan through 2020. The background (blue) carries a systematic −0.70 mm bias; the analysis that assimilates the station (orange) tracks the radiosonde (black dotted) closely, whereas the leave-one-out analysis at the same station (green dashed) remains near the background for the entire year (the green dashed line overlies the blue background because the two nearly coincide)

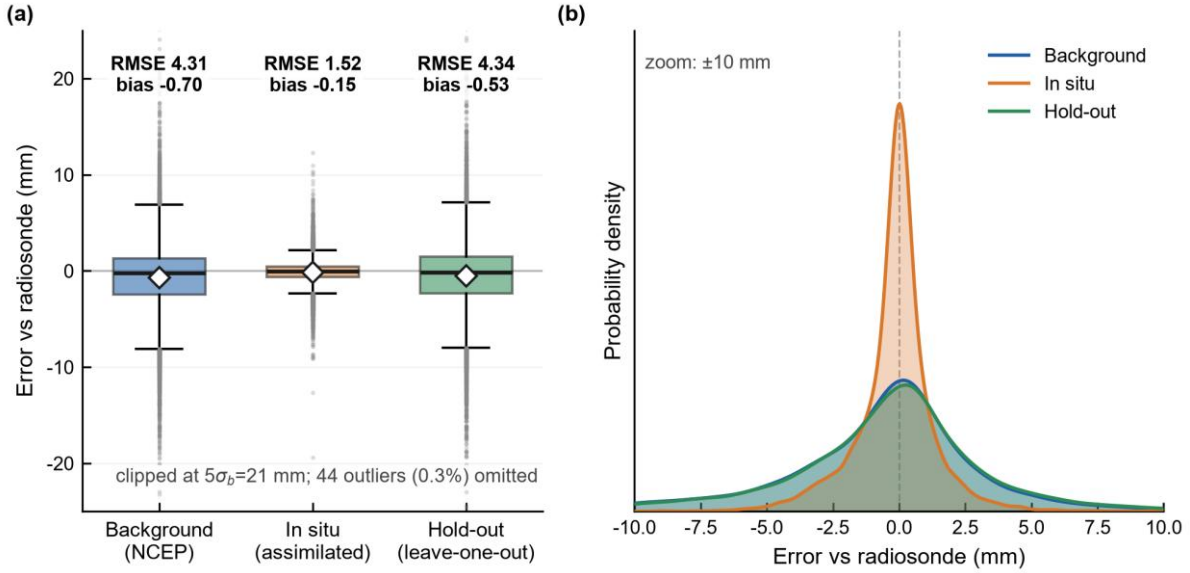


Fig. 3 Distributions of the per-station-day errors against radiosonde PWV over 14,256 station-days under the three scoring conventions, shown two ways. (a) Boxplot: median, IQR and 1.5×IQR whiskers, the diamond marking the mean. (b) Zero-centred (±10 mm; tails beyond this are truncated for display) Gaussian kernel-density estimate with Scott's bandwidth: the in-situ error collapses into a sharp spike at zero while the background and hold-out retain a broad spread and a residual bias. In both panels the in-situ errors collapse toward zero and lose their bias, whereas the hold-out errors retain both the spread and the bias of the background: a withheld location inherits the background error essentially unchanged. Both panels are clipped at the 5σ_b = 21 mm gross-error threshold (44 outliers, 0.3%, omitted for display); n = 14,256 and the RMSE/bias labels use the full sample

The gap is not an artefact of a few contaminated soundings. Applying the gross-error check that is standard practice before assimilation removes few data points and leaves the conclusion intact (Table 1): at 5σ_b only 22 of 14,256 station-days are rejected and the picture is unchanged (4.11 mm background, 1.47 mm in situ, 4.14 mm hold-out); even at 3σ_b, which rejects 234 station-days including implausible profiles such as one Kunming sounding whose self-integrated PWV of 72 mm lies far above the 5σ_b innovation threshold, the in-situ reduction stays near 64% and the hold-out change stays near −1%.

6.3 The gap holds at every station

Table 2 and Figure 4 decompose the result by station. The in-situ reduction is large everywhere but varies from +57.2% (Harbin) to +85.3% (Kunming), whereas the hold-out change is negligible everywhere, between −2.7% and +0.9%. The gap is therefore structural rather than station-specific. Its structure is revealing: Kunming both gains most in situ (+85.3%, leaving an RMSE of 0.42 mm) and gains nothing when withheld (+0.4%), so the

value-transfer gap is widest exactly where the background error is largest and a good analysis would be most valuable. Wuhan and Harbin are the only stations where the hold-out change is negative, and even there it does not exceed 2.7%.

Table 2 Per-station RMSE (mm) against radiosonde PWV, pooled over 2015–2022. "Change" is relative to that station's background RMSE; positive denotes an improvement (RMSE reduction). The distance column gives the separation from each station to its nearest neighbour in the network, which Section 6.5 shows to be the controlling quantity.

Station	n (station-days)	Background	In situ	Change	Hold-out	Change	Nearest neighbour (km)
Kunming	2,830	2.88	0.42	+85.3%	2.87	+0.4%	1,286 (Wuhan)
Beijing	2,860	5.29	1.68	+68.2%	5.24	+0.9%	1,053 (Wuhan)
Wuhan	2,863	4.54	1.59	+64.9%	4.66	−2.6%	704
Shanghai	2,872	4.77	1.90	+60.1%	4.75	+0.3%	704
Harbin	2,831	3.64	1.56	+57.2%	3.73	−2.7%	1,071 (Beijing)

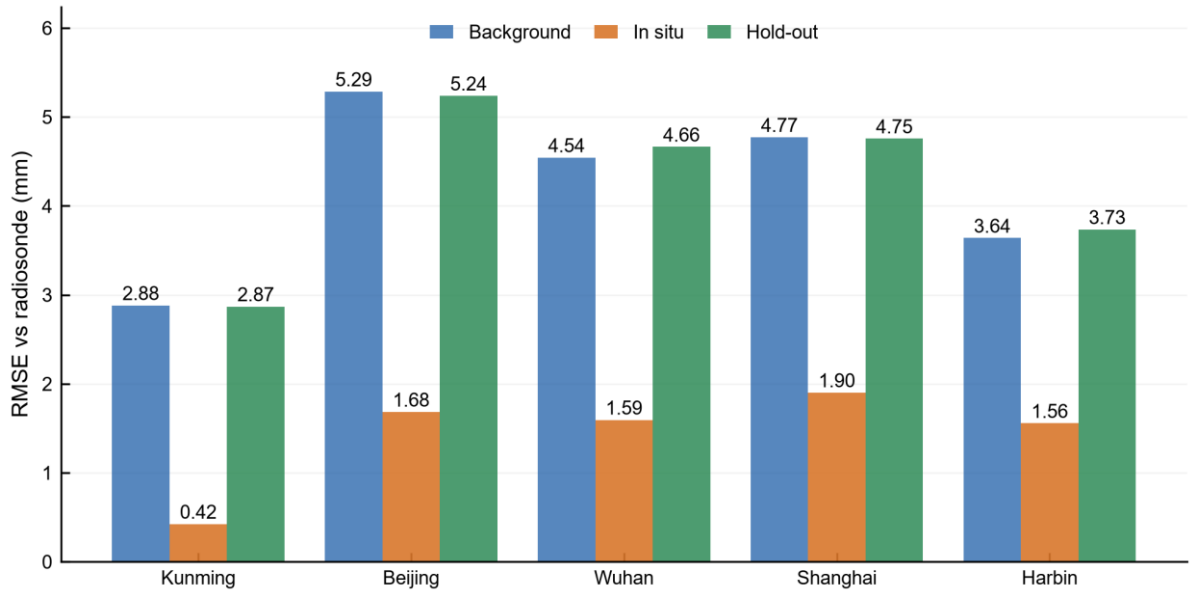


Fig. 4 Per-station RMSE against radiosonde PWV under the three scoring conventions. The in-situ RMSE is reduced by 57–85% at every station, while the hold-out RMSE remains within 3% of the background at every station, so the value-transfer gap repeats station by station rather than being an aggregate artefact

6.4 Robustness across years and seasons

A single year could in principle be unrepresentative, so the experiment was repeated for each of the eight years with that year's own σ_b. Figure 5 shows the gap to be persistent rather than incidental: the in-situ reduction stays between 62.5% and 66.0% in every year, while the hold-out change stays between −2.9% and +2.3% and never approaches the in-situ value.

The seasonal decomposition is more informative (Figure 6). In-situ skill is almost flat across the calendar year, from 62.1% in May to 69.1% in November, so the analysis is equally good whenever a sounding exists. The hold-out change, by contrast, varies systematically with the monsoon: it is negative in the dry season (−4.7% in January, when the mean PWV is 8.7 mm) and turns positive in the wet season (+3.2% in August, at 43.2 mm). The background error itself grows from 1.72 mm in January to 6.85 mm in July, yet the fraction of it that a distant network can remove grows faster still, so the transfer is least negative when the water-vapour field is dominated by the large-scale monsoon signal and most negative when small-scale dry-season variability dominates. This reinforces Section 6.5: what limits the transfer is whether the network resolves the scales that carry the signal.

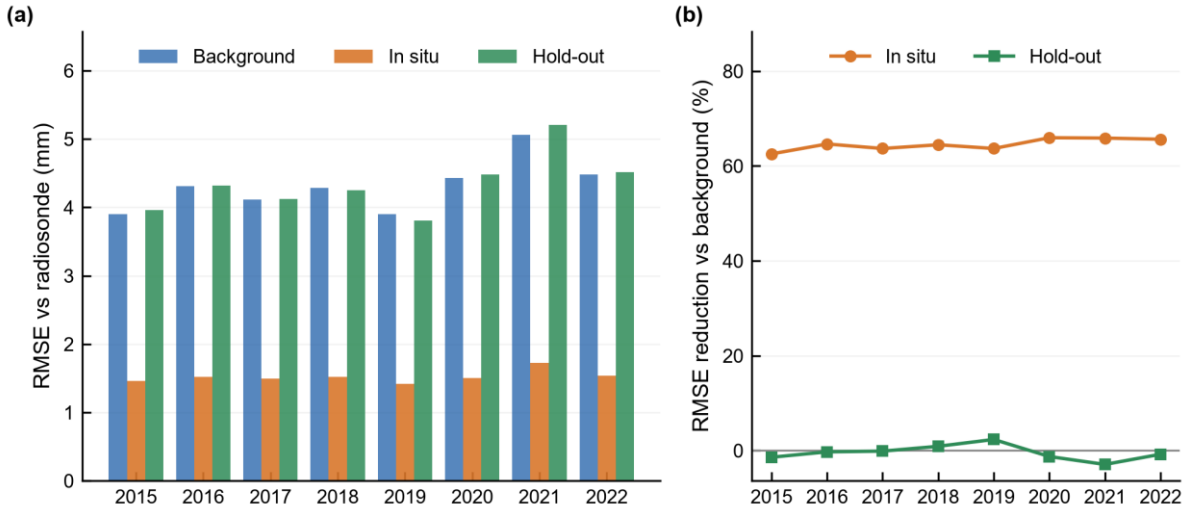


Fig. 5 Inter-annual robustness of the value-transfer gap, 2015–2022. (a) RMSE against radiosonde PWV for each year under the three scoring conventions. (b) The same results as the RMSE reduction relative to that year's background. The in-situ reduction stays between 62.5% and 66.0% in every year, whereas the hold-out change stays between −2.9% and +2.3%, so the gap is a property of the network rather than of one year's weather

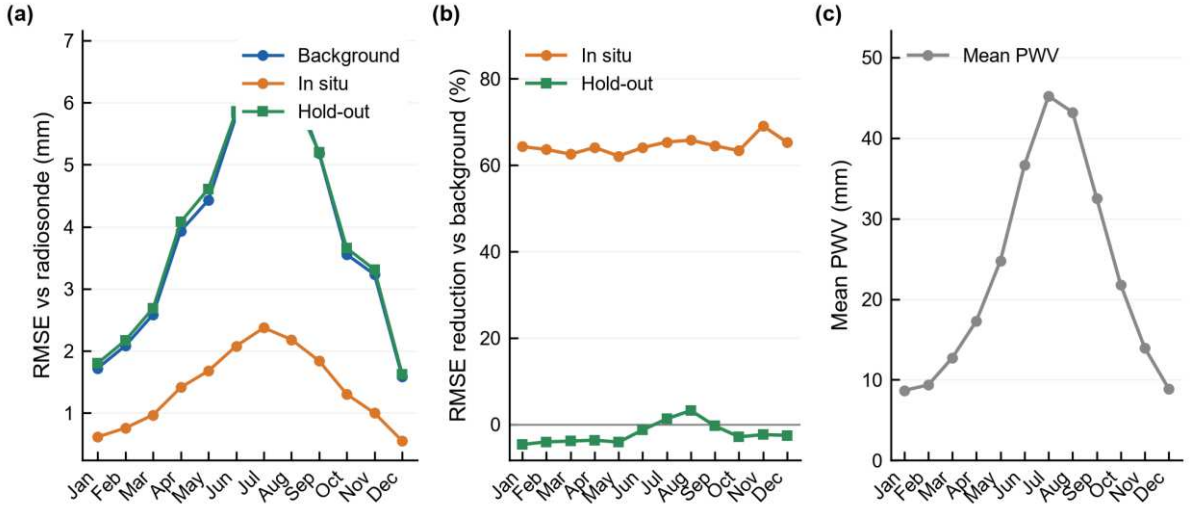


Fig. 6 Seasonal cycle of fusion skill, pooled over 2015–2022. (a) Monthly RMSE under the three scoring conventions. (b) The same results as the RMSE reduction relative to that month's background. (c) The background (NCEP) mean PWV, shown on its own axis as the monsoon-context reference for (a,b). In-situ skill is nearly flat across the year (62.1–69.1%), but the hold-out change is negative in the dry season (−4.7% in January) and positive in the monsoon months (+3.2% in August), so the value transfer depends on whether the water-vapour field is dominated by the large-scale monsoon signal or by small-scale dry-season variability

6.5 Why the analysis cannot be transferred: the decorrelation limit

The remaining question is whether the near-zero hold-out gain indicts the fusion method or the network, and the real data point to the network. Figure 7 plots the hold-out errors against the distance from each withheld station to the nearest assimilated station, together with the theoretical analysis-error standard deviation σ_a(d) implied by the same covariance used to build the analysis: σ_a(d) rises from 1.41 mm at zero separation (close to the observed in-situ RMSE of 1.52 mm) and returns to within a few percent of the background error by roughly 500 km. The static nearest-neighbour separations in this network are 704–1,286 km (mean ≈ 964 km) (Table 2), and on days when the nearest neighbour does not report, the effective separation to the nearest assimilated station is larger still, reaching ≈1,970 km in the binned display of Fig. 7 (and 2,087 km in the raw record, dist_nearest in fusion_multi.npz); even the closest operating radiosonde pair lies beyond the 400 km decorrelation length. The analysis-error standard deviation at separation d (8) is

$$\sigma_a^2(d) = \sigma_b^2 - \frac{\sigma_b^4}{\sigma_b^2 + \sigma_o^2} exp(-\frac{2d}{L}) \quad (8)$$

which gives 1.41 mm at d = 0 and returns to within a few percent of the background by roughly 500 km. Under those conditions the optimal-interpolation update at a withheld site is a small fraction of the background error, and the hold-out analysis reduces to the background by construction.

The near-zero hold-out gain is therefore not a failure of the fusion method; it is the correct behavior of a well-specified analysis applied to a network whose spacing exceeds its own correlation length. This reframes the practical question from which fusion algorithm is best to whether a given network resolves the scales at which fusion can transfer value, and it makes network density (not algorithm choice) the master variable taken up in Section 7. It also resolves the apparent tension noted in Section 6.8: dense networks such as CMONOC (Ren et al. 2022) do report 10–15% hold-out gains precisely because their station spacing lies inside the decorrelation length, so the two results are consistent once density is stated explicitly. That the same covariance predicts value transfer inside L and saturation outside it is shown directly in the controlled leave-one-out experiment of Figure 9, which uses no real station at all and therefore cannot be accused of selection.

Across the five sites the normalized ratio reaches 0.33 at zero separation, a 67% reduction relative to the background, whereas every observed bin lies on the saturated tail at 0.99–1.03; the two rightmost bins (n = 50 and 36) are days on which the nearest neighbour did not report. Normalizing by each station's own background removes the 2.3 mm spread in raw background quality across sites (r = 0.997 between a station's background and hold-out RMS), so the ratio isolates the distance effect.

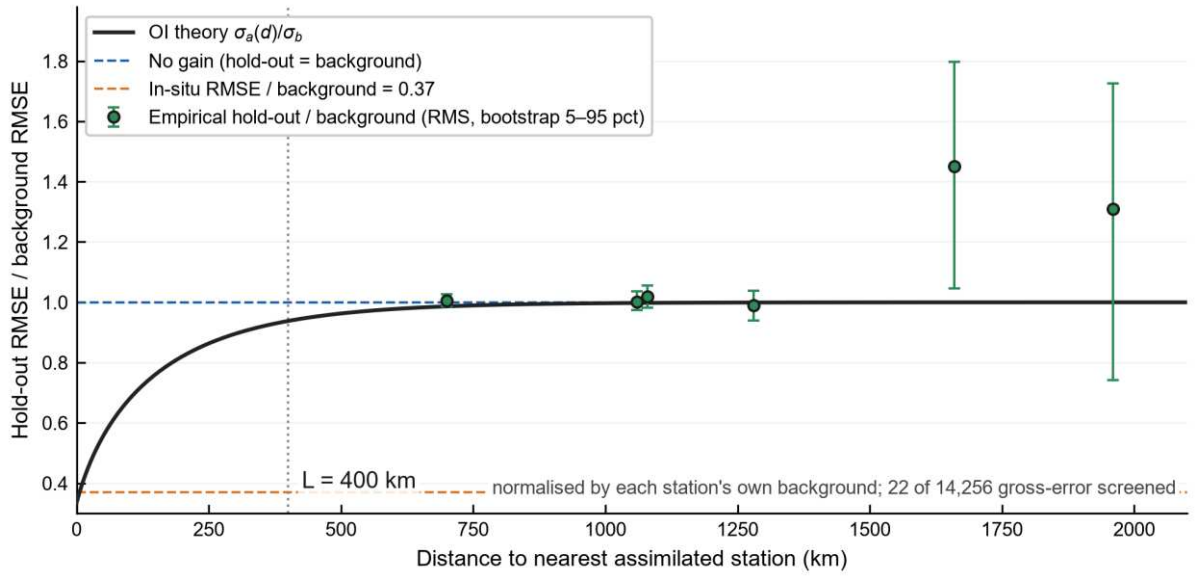


Fig. 7 Value transfer versus station separation. Green points are leave-one-out RMS errors normalized by each station's own background RMS, binned to the nearest 20 km (bins with ≥30 station-days; 14,229 screened station-days in total) against distance to the nearest assimilated station; whiskers give the 5–95 percentile interval over 200 bootstrap resamples. The black curve is the single-station theoretical optimal-interpolation error of Eq. (8) in normalized form, σ_a(d)/σ_b (σ_b = 4.22 mm, σ_o = 1.5 mm, L = 400 km). The dashed blue line is the no-gain reference (hold-out = background) and the dashed orange line the in-situ ratio 1.52/4.11 = 0.37

6.6 A real forecasting benchmark that fails to beat persistence

The second half of the claim concerns forecasting, and here we report a deliberately modest but real benchmark because its outcome is instructive. A minimal LSTM (Hochreiter and Schmidhuber 1997) was trained on the Wuhan IGRA-v2 PWV series for 2015–2022 with a 14-day input window and a direct 7-step output head, holding out the 2022 year, with persistence as the baseline. The model was kept small (16 hidden units, 90 epochs, and 830 of the 2,491 available training windows) so this is a benchmark rather than a tuned system.

Persistence is hard to beat at daily resolution, and this model does not beat it: over the 2022 test year the LSTM RMSE is 16.11 mm against 12.34 mm for persistence, 30.5% worse (Table 3; Figure 8a). The LSTM error is nearly flat across leads (15.6–16.4 mm from 1 to 7 days) whereas persistence degrades gradually from 7.63 to 14.50 mm, so the two curves converge with lead but never cross. Figures 8b and 8c show the mechanism: the LSTM has collapsed toward the series mean and reproduces little of the day-to-day variability, while persistence inherits it by construction. We state the limits of this negative result plainly (one station, one minimal architecture, a subsampled training set) so it does not establish that deep learning cannot beat persistence on PWV. What it does illustrate, consistently with the mixed evidence of Section 4, is that a forecasting gain is not automatic: it depends strongly on lead time, on which baseline is chosen, and on whether the model is given skill-carrying predictors at all. The persistence baseline (9) and the minimal LSTM cell (10) are

$$\hat{x}(t+\tau) = x(t) \tag{9}$$

$$f_t = \sigma_g(W_f[h_{t-1}, x_t] + b_f), i_t = \sigma_g(W_i[h_{t-1}, x_t] + b_i), c_t = f_t \odot c_{t-1} + i_t \odot \tilde{} c_t, h_t = o_t \odot tanh(c_t) \quad (10)$$

Table 3 Short-term PWV forecasting at Wuhan (IGRA-v2 radiosonde, 2015–2022; 2022 held out). RMSE and MAE in mm for the minimal LSTM and the persistence baseline by forecast horizon. The final row pools all horizons.

Lead (days)	LSTM RMSE	Persistence RMSE	LSTM MAE
1	15.65	7.63	13.30
2	15.92	10.36	13.52
3	16.03	11.80	13.62
4	16.18	13.01	13.73
5	16.25	13.66	13.82
6	16.33	13.97	13.89
7	16.37	14.50	13.95
All	16.11	12.34	—

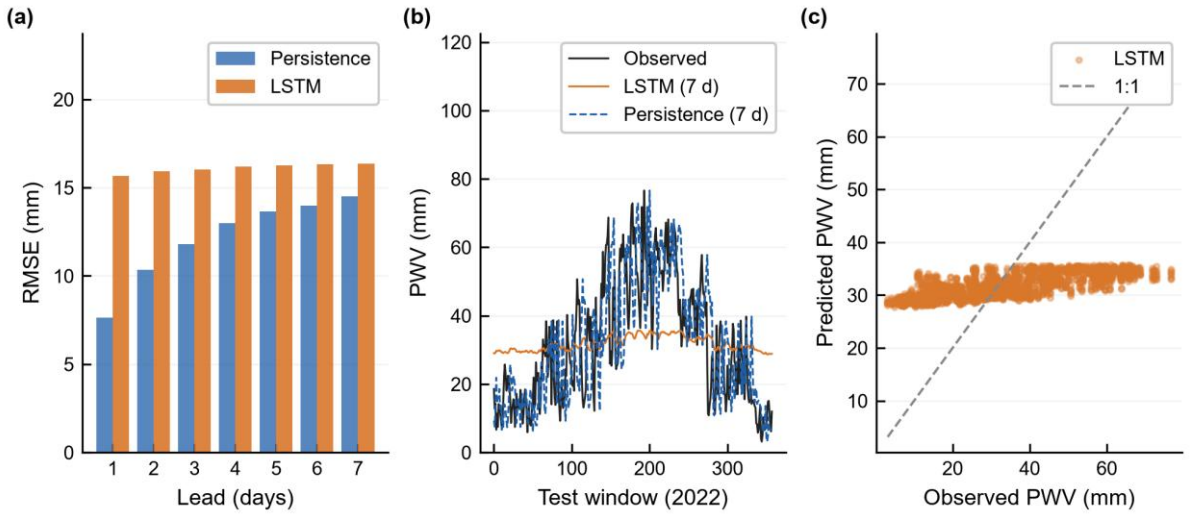


Fig. 8 Real-data short-term forecasting benchmark at Wuhan. (a) RMSE by forecast horizon for the minimal LSTM (orange) and persistence (blue). (b) Test-year (2022) time series of observed PWV (black), the LSTM forecast at 7-day lead (orange) and persistence at the same lead (blue dashed): the LSTM has collapsed toward the mean and tracks little of the variability. (c) LSTM predictions against observations for all horizons pooled (orange)

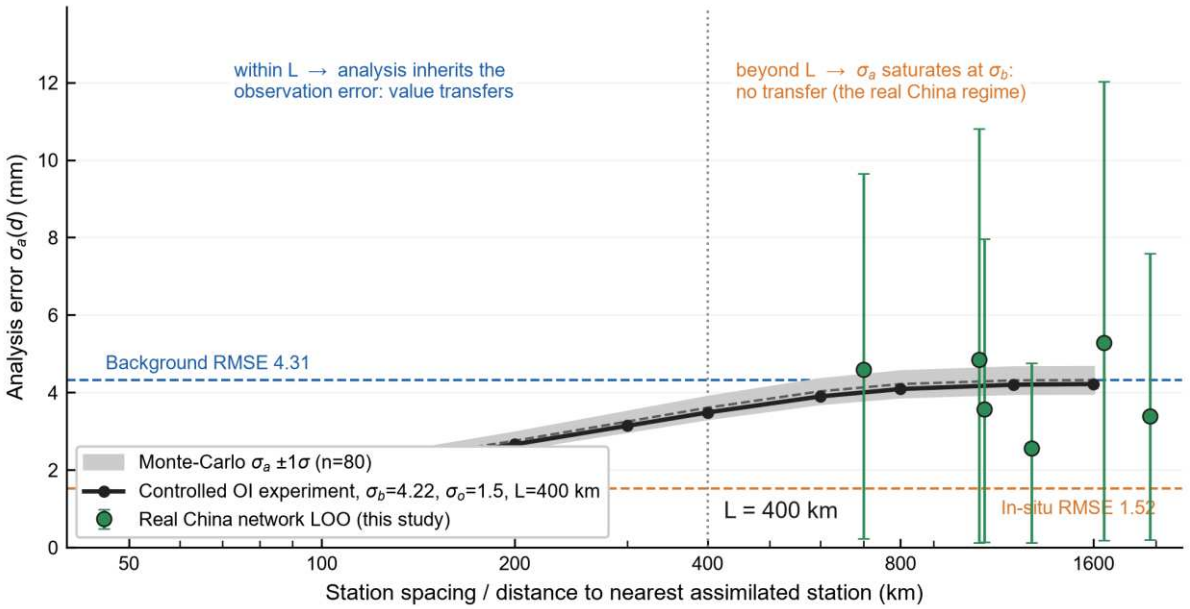


Fig. 9 Controlled optimal-interpolation experiment isolating network spacing. Leave-one-out OI analyses on synthetic regular lattices of spacing h use the exponential covariance of Section 6.5 (σ_b = 4.22 mm, σ_o = 1.5 mm, L = 400 km). The black curve is the deterministic lattice result, which lies below the single-station curve of Eq. (8) because several neighbours are assimilated at once; the grey band is ±1 standard error of the Monte-Carlo estimate (80 draws; SE ≈ 0.3 mm). Dashed lines are the in-situ (1.52 mm) and background RMSE references, which include residual bias (RMSE² = σ² + bias²) and

therefore sit slightly above the σ_a(0) limit of 1.41 mm. Green points with 5–95 percentile bootstrap whiskers are the real China-network leave-one-out RMS errors

6.7 Does the analysis gain transfer to a forecast? An initial-condition test

The forecasting half of the value-transfer question has so far rested on the literature, because the 6.6 benchmark never fed a fused field into a forecaster. We close that gap on our own data. For every station-day we initialise an identical damped-persistence forecaster—$fc(k) = c + \rho^{y}(IC(t) - c)$, with c the station climatology and ρ its lag-1 autocorrelation—from three initial-condition (IC) fields at day t: the background (no fusion), the OI analysis that assimilated the station (dense case), and the OI analysis with the station withheld (sparse case). The forecast is verified against the independent radiosonde at day t+k (k = 1…7 d). The verification target is never used to build the IC, so there is no leakage, and a block bootstrap over (station, year, month) cells supplies 95% confidence intervals that respect serial correlation.

In dense regions the fused-initialised forecast is significantly better at every lead (Table 4; Fig. 10): lead-1 RMSE falls by 0.38 mm (95% CI [−0.46, −0.30] mm, about 5% of the forecast error and 13.6% of the 2.79 mm analysis-domain gain), and the transfer efficiency—the fraction of that gain which survives into the forecast—decays only from 13.6% at day 1 to 9.4% at day 7. In sparse regions the fused IC equals the background and no transfer is detected (all 95% CIs include zero). A trajectory-based multilayer-perceptron forecaster trained on a 14-day window did not resolve a significant transfer; its weaker skill (RMSE ≈ 9 mm) leaves it underpowered for the sub-0.4 mm effect, and more fundamentally the gain from a single improved initial value contributes little once forecast skill is dominated by temporal evolution. The transfer is therefore real but forecaster-structure dependent: sizable for initial-condition-driven forecasting, negligible for evolution-dominated forecasting.

Lead (d)	RMSE bg	RMSE insitu	RMSE holdout	in−bg (mm)	95% CI	TE in (%)
1	7.215	6.836	7.209	−0.379	[−0.459, −0.298]	13.6
2	9.264	8.921	9.238	−0.345	[−0.411, −0.273]	12.3
3	10.021	9.699	9.987	−0.322	[−0.386, −0.253]	11.6
4	10.426	10.114	10.383	−0.311	[−0.370, −0.245]	11.2
5	10.802	10.504	10.756	−0.298	[−0.356, −0.239]	10.7
6	11.214	10.937	11.169	−0.277	[−0.333, −0.219]	9.9
7	11.654	11.391	11.609	−0.263	[−0.311, −0.207]	9.4

Table 4 Value transfer from analysis to forecast, by lead, from the identical damped-persistence forecaster initialised from the background (bg), the in-situ OI analysis (dense), and the hold-out OI analysis (sparse). RMSE in mm against the independent radiosonde. in−bg is the dense transfer (negative = fused better); TE_in is the transfer efficiency as a percentage of the analysis-domain gain (2.79 mm). All dense 95% CIs exclude zero; all sparse CIs include zero.

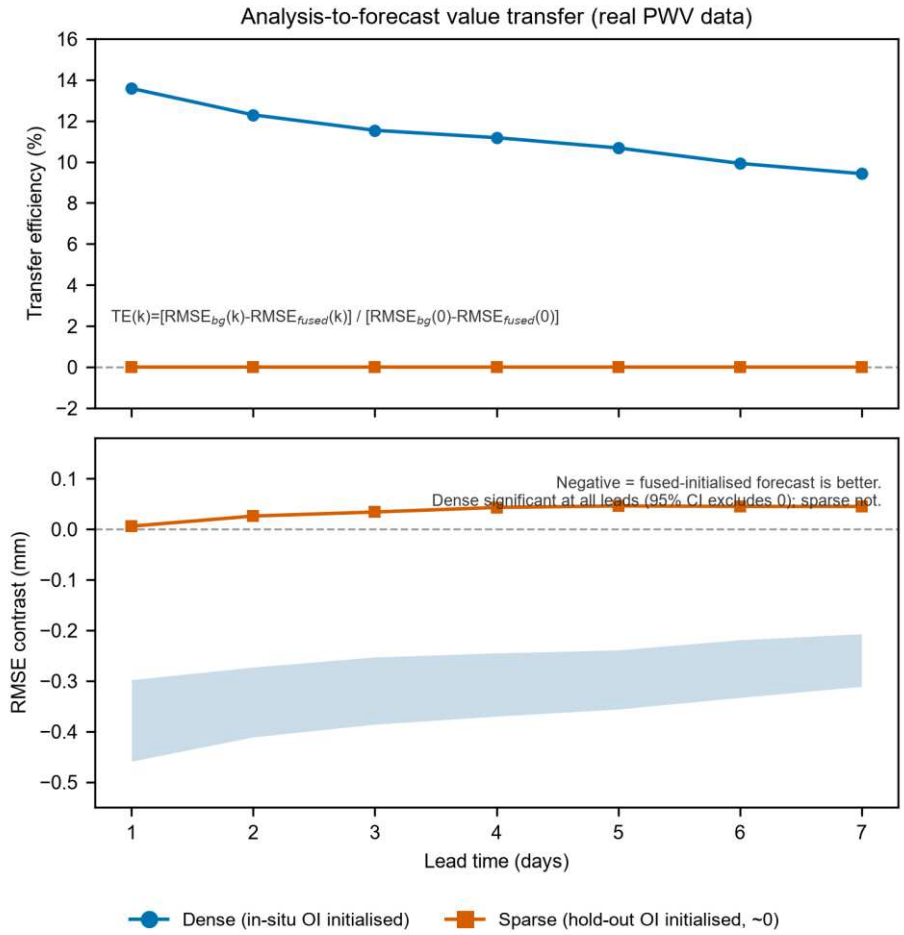


Fig. 10 Direct test of the analysis-to-forecast value transfer on real multisource PWV data. (a) Transfer efficiency—the share of the analysis-domain RMSE gain that survives into a k-day forecast—against lead time for the dense (in-situ initialised) and sparse (hold-out initialised) cases; the dense band is the 95% bootstrap CI and excludes zero at every lead. (b) Absolute RMSE contrast (fused minus background initialised) in mm; negative means the fused-initialised forecast is better, which is significant only in the dense case. The figure quantifies the gap: roughly one tenth of the analysis improvement reaches a 1–7 d forecast in dense networks and essentially none in sparse ones

6.8 What the demonstration shows and where it stops

This demonstration uses genuinely observed PWV throughout: the observations are total-column PWV self-integrated from raw IGRA-v2 soundings, and both background and analysis are verified against those same soundings, so no reanalysis value is ever used as ground truth.

What it shows is threefold. First, the value-transfer gap is quantified and mechanistically explained on real data: fusion is decisive where observations are dense (+64.7% in situ) and near-null where they are not (−0.6% at held-out sites), and it follows from a network spacing that exceeds the analysis correlation length. Second, that gap is robust to standard gross-error screening, so it is not an artefact of contaminated soundings. Third, and only illustratively, a forecasting gain is not automatic: on real daily PWV a minimal deep-learning model fails to beat persistence at any horizon up to seven days, which is consistent with the mixed evidence of Section 4 rather than a proof about deep learning. The 6.7 initial-condition test goes further: when a forecaster is initialised from the fused field the gain does transfer in dense networks and not in sparse ones.

Three limits bound the demonstration. First, the daily 2.5° background and the sparse five-station network make the absolute numbers (4.31 mm background, 1.52 mm in situ) a deliberately pessimistic lower bound rather than typical performance, so they are not comparable to the sub-daily, 0.25° or dense-GNSS results of Sections 3 and 4; likewise the −0.6% is the sparse-network limit, consistent with rather than contradictory to the 10–15% hold-out gains of dense networks once density is stated. Second, and following directly from Section 6.5, the near-zero hold-out gain is a property of this network geometry and is not a bound on what a dense network would achieve. Third, the 6.7 initial-condition test now supplies the study's own direct evidence for the forecasting-transfer claim—dense networks transfer about 10–14% of the analysis gain into 1–7 d forecasts and sparse networks none—so the burden that remains is for dense-network and deep-learning studies such as Hu et al. (2026), whose fused ZTD grid fed into WRF cuts PWV forecast error by about 30%, to meet the hold-out,

direct-GNSS, multi-seed verification demanded by the Section 5 checklist, rather than to establish that transfer exists.

7 Challenges and future directions

Several obstacles stand between the current literature and operational, trustworthy PWV products.

Network density remains the master variable: fusion and forecasting collapse to the background where observations are absent, and assimilating continental CORS networks (CMONOC, provincial networks) directly is the natural next step where they are not yet in the processing chain.

Evaluation rigor must become standard. Hold-out and direct-GNSS verification, multi-seed block bootstrap, and an explicit ERA5-only control should be the default, not the exception.

Real-time latency is a practical barrier—ERA5 itself carries about a five-day delay, so operational products must predict the reanalysis or use faster streams (Meng et al. 2026 predict ERA5 PWV to close this gap).

Extreme-weather focus is under-developed: most skill is reported as domain mean RMSE, which hides the strong-gradient, high-impact cases where fusion should matter most and where a forecast error is most costly. Targeted validation on heavy-rain, typhoon and frontal cases (not domain means) is what matters for operations, and is exactly where the hold-out and direct-GNSS conventions of Figure 1 are hardest to satisfy and most informative.

Multi-source richness is increasing (FY-4A, FY-3D, InSAR, MODIS all now fused with GNSS and ERA5; Guo et al. 2026 combine four sources over the Jiaodong Peninsula), but systematic comparison of which source adds what, in which regime, is missing.

Uncertainty quantification is nearly absent: fusion and forecasting both report point RMSE with little on the error distribution or on where the model is wrong.

Finally, foundation models and data-driven weather models (e.g., the WeatherBench2 benchmark, Rasp et al. 2024) are beginning to absorb PWV as one variable among many; how fusion and station-level forecasting plug into them is an open and promising direction, and it reframes the fusion question from "build a better PWV field" to "supply the surface data-assimilation step that these models still lack".

Two obstacles stand out because they are tractable now: Uncertainty quantification is not merely a reporting nicety: without prediction intervals a forecaster cannot tell a 1 mm confident forecast from a 5 mm uncertain one, and the fusion community already has the tools (variance-component estimation (Du et al. 2025), ensemble ML, and the Desroziers diagnostic) to produce them. Cross-source attribution is the other: as FY-4A, FY-3D, InSAR and MODIS are fused together (Guo et al. 2026), the field needs ablation studies that add sources one at a time so that the marginal value of each, in each regime, becomes known rather than assumed. Closing both gaps is what would make the current scattered results operational.

8 Conclusions

Multi-source PWV fusion and deep-learning short-term forecasting have matured into two substantial, if still parallel, literatures. Fusion now turns a discontinuous, accurate observing network and a biased but complete reanalysis into a spatially complete field with analysis RMSE of roughly 0.6–3.3 mm against GNSS and hold-out gains of 10–15%; deep learning now forecasts PWV from station histories or spatial fields with RMSE of 0.8–8.1 mm depending on network density and terrain. The field's outstanding problem is not method but evidence: the claim that fusion improves forecasting is real in dense, well-observed regimes and when the forecast model consumes the fused field, but it is protocol-dependent and frequently unverified, and it has not been shown to survive averaging over seeds or sparse networks. For practitioners, the actionable message is narrow but firm: trust a fused field for analysis, but do not assume it improves a forecast unless that forecast was verified against independent GNSS under an ERA5-only control and across seeds. The reporting checklist in Section 5 is a minimal bar for that verification, and Figure 1 is the visual shorthand for why the distinction matters. On our own data the fused field does improve a 1–7 d forecast by about 0.4 mm (≈14% of the analysis gain) in dense networks and not at all in sparse ones (Section 6.7), so the qualification is density- and forecaster-dependent rather than absolute.

We recommend that future work report hold-out fusion skill, direct GNSS forecast verification, multi-seed block-bootstrap significance, and an explicit ERA5-only control, and design fusion and forecasting as one pipeline rather than two. Doing so would turn today's scattered, individually impressive results into a trustworthy operational capability.

Declarations

Funding. This research received no specific grant from any funding agency.

Competing interests. The author declares no competing interests.

Data and code availability. The empirical demonstration of Section 6 uses publicly archived data: 2015–2022 daily NCEP/NCAR Reanalysis precipitable water (pr_wtr; Kalnay et al. 1996) from the NOAA Physical

Sciences Laboratory, and IGRA-v2 radiosonde soundings. The fusion code and the exact data identifiers are available from the author upon reasonable request.

Author contributions. D.D. conceived the study, conducted the literature review and the empirical analysis, and wrote the manuscript. D.D. read and approved the final version.

Generative AI statement. The author used a large-language-model assistant for language editing and literature organization; all factual and numerical claims were verified against the primary sources cited in this paper.

Askne J, Nordius H (1987) Estimation of tropospheric delay for microwaves from surface weather data. *Radio Sci* 22(3):379–386. https://doi.org/10.1029/RS022i003p00379

Bevis M, Businger S, Herring TA, Rocken C, Anthes RA, Ware RH (1992) GPS meteorology: remote sensing of atmospheric water vapor using the global positioning system. *J Geophys Res* 97(D14):15787–15801. https://doi.org/10.1029/92JD01517

Blewitt G, Hammond WC, Kreemer C (2018) Harnessing the GPS data explosion for interdisciplinary science. *Eos* 99. https://doi.org/10.1029/2018EO104623

Cao Y, Tang J, Li H, Yao Y, Zhang L, Xu C (2026) Enhancing short-term PWV prediction through GNSS and ERA5 data fusion. *Atmos Res* 331:108663. https://doi.org/10.1016/j.atmosres.2025.108663

Chang C, Liu Y, Wang Y, Zhang Y, Meng X, Chi H, He C (2026) Railway CORS GNSS PWV forecasting using normal time-frequency transform and deep learning with transfer learning strategy. *Adv Space Res* 77(7):7749–7767. https://doi.org/10.1016/j.asr.2026.02.018

Chen G, Xu A, Xu Z, Deng Z, Tang L, Shao C, Yang N, Gao X, Zhang M (2026) Synergizing GNSS, MODIS, and ERA5 for high-resolution PWV retrieval: a two-stage machine learning approach over Hong Kong. *Adv Space Res* 77(6):6607–6628. https://doi.org/10.1016/j.asr.2026.01.035

Daley R (1991) *Atmospheric data analysis*. Cambridge University Press, New York.

Desroziers G, Berre L, Chapnik B, Poli P (2005) Diagnosis of observation, background and analysis-error statistics in observation space. *Q J R Meteorol Soc* 131(613):3385–3396. https://doi.org/10.1256/qj.05.108

Ding M, Ding J, Peng Z, Su M, Sun T (2025) Developments of empirical models for vertical adjustment of precipitable water vapor measured by GNSS. *Adv Space Res*75(3):2473–2483. https://doi.org/10.1016/j.asr.2024.08.039

Duan J, Bevis M, Fang P, et al. (1996) GPS meteorology: direct estimation of the absolute value of precipitable water. *J Appl Meteorol* 35(6):830–838. https://doi.org/10.1175/1520-0450(1996)035<0830:GMDEOT>2.0.CO;2

Du W, Wang Y, Meng X, Liu Y, Liu X (2025) Research on the fusion of PWV over the Beijing-Tianjin-Hebei region based on ERA5 PWV and FY-4A PWV. *Adv Space Res* 75(7):5311–5330. https://doi.org/10.1016/j.asr.2025.01.025

Emardson TR, Elgered G, Johansson JM (1998) Three months of continuous monitoring of atmospheric water vapor with a network of Global Positioning System receivers. *J Geophys Res* 103(D2):1807–1820. https://doi.org/10.1029/97JD03015

Gandin LS (1963) *Objective analysis of meteorological fields*. Israel Program for Scientific Translations, Jerusalem.

Ghaffari-Razin SR, Davari Majd R, Hooshangi N (2023) Regional modeling and forecasting of precipitable water vapor using least square support vector regression. *Adv Space Res* 71(11):4725–4738. https://doi.org/10.1016/j.asr.2023.01.030

Guerova G, Jones J, Douša J, et al. (2016) Review of the state of the art and future prospects of the ground-based GNSS meteorology in Europe. *Atmos Meas Tech* 9(11):5385–5406. https://doi.org/10.5194/amt-9-5385-2016

Guo Q, Wang Z, Li D, Sun Y, Zhang J, Liu H (2026) Construction of a continuous high-resolution PWV using GNSS/ERA5, InSAR, and FY-4A data: a case study of the Jiaodong Peninsula and adjacent seas. *Appl Sci* 16(11):5391. https://doi.org/10.3390/app16115391

Haji-Aghajany S, Izanlou S, Tasan M, Rohm W, Kryza M (2025) High-resolution GNSS troposphere tomography through explainable deep learning-based downscaling framework. *Satellite Navigation* 6:22. https://doi.org/10.1186/s43020-025-00177-6

Hersbach H, Bell B, Berrisford P, et al. (2020) The ERA5 global reanalysis. *Q J R Meteorol Soc* 146(730):1999–2049. https://doi.org/10.1002/qj.3803

Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu F, Sha Z, Wei P, Xia P, Ye S, Zhu Y, Luo J (2024) Deep learning for GNSS zenith tropospheric delay forecasting based on the Informer model using 11-year ERA5 reanalysis data. *GPS Solut* 28(4):182. https://doi.org/10.1007/s10291-024-01720-9

Hu F, Wei P, Sha Z, Ye S, Xia P, Zhao L, Luo J (2026) A fusion ZTD grid model utilizing ERA5 and GNSS raw observations for enhancing PPP and optimizing PWV forecasts under extreme weather conditions. *GPS Solut* 30(2):65. https://doi.org/10.1007/s10291-026-02028-6

Kalnay E, Kanamitsu M, Kistler R, et al. (1996) The NCEP/NCAR 40-year reanalysis project. *Bull Am Meteorol Soc* 77(3):437–472. https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2

Li H, Zhao Q, Guo H, Li Z, Ma Y, Yao Y, Yin J, Zhai Y, Liang H, Xiong Z (2026) A GNSS PWV filling and short-term forecasting framework fused hybrid neural network. *Atmos Res* 329:108508. https://doi.org/10.1016/j.atmosres.2025.108508

Li J, Li F, Liu L, Yao Y, Huang L, Wang Y (2024) A weighted mean temperature forecast model based on fused data and generalized regression neural network and its impact on GNSS-based PWV estimation. *IEEE Trans Geosci Remote Sens* 62:1–14. https://doi.org/10.1109/TGRS.2024.3353300

Lin Z, Li M, Zheng Z, Cheng Y, Yuan C (2020) Self-attention ConvLSTM for spatiotemporal prediction. *Proc AAAI Conf Artif Intell* 34(7):11531–11538. https://doi.org/10.1609/aaai.v34i07.6819

Ma X, Huang Y, Zhang Q, Zhang B, Gao X, Zhao Q, Lin X, Xu C, Yao Y (2026) PWVFnet: a short-time troposphere PWV forecast model combining empirical models and spatiotemporal ConvLSTM network. *IEEE J Sel Top Appl Earth Obs Remote Sens* 19:3858–3871. https://doi.org/10.1109/JSTARS.2025.3648114

Meng X, Wang Y, Zhang Y, Du W, Liu Y, Liu X (2026) Quasi-real-time retrieval of ERA5 precipitable water vapor over mainland China. *Adv Space Res* 77(6):6952–6975. https://doi.org/10.1016/j.asr.2026.01.064

Pang H, Liu W, Wang Y, Huang L, Zhang L, Wang X (2024) A GRNN-based model for ERA5 PWV adjustment with GNSS observations considering seasonal and geographic variations. *Remote Sens* 16(13):2424. https://doi.org/10.3390/rs16132424

Rasp S, Hoyer S, Merose A, et al. (2024) WeatherBench 2: a benchmark for the next generation of data-driven global weather models. *J Adv Model Earth Syst* 16(6):e2023MS004019. https://doi.org/10.1029/2023MS004019

Ren D, Wang Y, Wang G, Liu L (2022) Fusion of CMONOC and ERA5 PWV products based on backpropagation neural network. *Remote Sens* 14(15):3750. https://doi.org/10.3390/rs14153750

Rocken C, Ware R, Van Hove T, Solheim F, Alber C, Johnson J, Bevis M, Businger S (1993) Sensing atmospheric water vapor with the global positioning system. *Geophys Res Lett* 20(23):2631–2634. https://doi.org/10.1029/93GL02935

Shi X, Chen Z, Wang H, Yeung D-Y, Wong W-K, Woo W-C (2015) Convolutional LSTM network: a machine learning approach for precipitation nowcasting. *Adv Neural Inf Process Syst* 28:802–810.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. *Adv Neural Inf Process Syst* 30:5998–6008.

Wijaya DD, Putri NSE, Utama AK, Wibowo ST, Sadarviana V (2024) Determination and assessment of GNSS-derived precipitable water vapor in Indonesia using Ina-CORS. *Adv Space Res* 73(1):386–403. https://doi.org/10.1016/j.asr.2023.07.048

Yang H, Ferreira V, He X, Zhan W, Wang X, Ji S (2025) Spatially enhanced interpolating vertical adjustment model for precipitable water vapor. *J Geod* 99(2). https://doi.org/10.1007/s00190-025-01936-8

Zhu C, Chen P, Yao Y, Qiu L, Zhang C, Lu J, Zhang Y, Yang X, Wu M, Zhang S (2024) High-resolution reconstruction and correction of FY-4A precipitable water vapor in China using back propagation neural network. *Atmos Res* 304:107357. https://doi.org/10.1016/j.atmosres.2024.107357

Zhu D, Li W, Zhang K, Hu Q, He P, Zhang L, Wu S, Yin W, Gao M, Li L (2026) A multichannel CNN-LSTM-based prediction model for precipitable water vapor in a region with a single GNSS station. *IEEE J Sel Top Appl Earth Obs Remote Sens* 19:3502–3514. https://doi.org/10.1109/JSTARS.2025.3649502

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- submission.zip