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.

Climate-to-Storage: Leakage-Controlled ConvLSTM Forecasting of GRACE-Derived Terrestrial Water Storage over Tamil Nadu, India

Rohit P
Amrita School of Engineering, Amrita Vishwa Vidyapeetham
Amirthalakshmi T M
tm_amirthalakshmi@ch.amrita.edu

Amrita School of Engineering, Amrita Vishwa Vidyapeetham
Rahul S G
Amrita School of Engineering, Amrita Vishwa Vidyapeetham

Article

Keywords: GRACE, GRACE-FO, terrestrial water storage, ConvLSTM, TerraClimate, data leakage, persistence baseline, skill score, drought index, Tamil Nadu

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11340997/v1

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

Additional Declarations: No competing interests reported.

Climate-to-Storage: Leakage-Controlled ConvLSTM Forecasting of GRACE-Derived Terrestrial Water Storage over Tamil Nadu, India

Rohit P [1], Amirthalakshmi T. M. [1,*] and Rahul S. G. [1]

[1] Department of Electronics and Communication Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Chennai, India; rohitpalanisamy608@gmail.com (R.P.); sg_rahul@ch.amrita.edu (R.S.G.)

[*] Correspondence: tm_amirthalakshmi@ch.amrita.edu (A.T.M.)

Abstract: Deep networks are increasingly used to forecast satellite gravimetry, but their apparent skill depends on how much memory of the storage record they receive. We examine this for Tamil Nadu using GRACE and GRACE Follow-On mascon terrestrial water storage (238 months, 2002-2024) and five TerraClimate variables. Each field was dated from its observation interval; count-based labelling would have misdated 81 of 238 fields. All masks and statistics came from training months only; 14 automated checks guarded against leakage. A convolutional long short-term memory (ConvLSTM) network predicted the next month from 18 months of climate alone (Model B) or with the latest storage field added (Model A). Over 19 test months, Model B had a root-mean-square error (RMSE) of 7.84 cm, against 12.76 cm for climatology and 12.26 cm for a Random Forest, yet persistence was more accurate (5.12 cm) and anomaly correlation only 0.365. About 94% of Model B's gain over climatology came from correcting bias in a wet period. Model A reached 4.52 cm but was not significantly better than persistence ($p = 0.374$; low power). At this lead and in this region, climate forcing therefore adds limited month-ahead information, and persistence should always be reported as a baseline.

Keywords: GRACE; GRACE-FO; terrestrial water storage; ConvLSTM; TerraClimate; data leakage; persistence baseline; skill score; drought index; Tamil Nadu

Introduction

Water held on and beneath the land surface, in aquifers, soils, rivers, reservoirs and irrigation tanks, controls how a region weathers a failed monsoon. The aggregate of these stores, terrestrial water storage, has been observable from orbit since 2002, when a pair of co-orbiting spacecraft of the Gravity Recovery and Climate Experiment (GRACE) started tracking how the planet's gravity varies from one month to the next.[1] GRACE Follow-On (GRACE-FO) resumed the measurements in 2018 after an eleven-month interruption.[2] Over land, most of the monthly gravity change is caused by water moving between places, so the missions deliver a direct, if coarse, measure of how much water a region has gained or lost.[3]

Indian applications of these data have been influential. Depletion of aquifers beneath the north-western plains was first quantified from GRACE,[4] and later work showed that storage in the south and east rises and falls with monsoon rainfall while the north continues to decline under pumping.[5] Global syntheses have placed these regional signals within a wider pattern of

emerging freshwater stress[6,7] and have shown that hydrological models often understate the trends that GRACE records.[8]

Three properties of the record limit its direct use in water management. Mascon solutions are distributed on fine grids, yet all cells inside one three-degree cap share a single value, which leaves few genuinely independent locations.[9,10] The series has isolated missing months and an inter-mission gap. Finally, each field is released only after processing ends, with a delay of some months, so the most recent fields are not available at the time a forecast is needed. These limits have motivated a growing body of work that predicts or reconstructs storage from climate variables[11,12] or with deep networks.[13–15]

Statistical reconstructions driven by precipitation and temperature reproduce much of the subdecadal variability of storage but cannot recover trends caused by human withdrawals.[11,12] Convolutional networks have been used to learn the mismatch between a land-surface model and GRACE over India,[13] to bridge the inter-mission gap with uncertainty estimates,[14] and to downscale storage for drought monitoring.[15] Recurrent models trained on meteorological forcing have matched calibrated conceptual models for runoff,[16] building on the gated memory of the long short-term memory (LSTM) network.[17] The convolutional LSTM (ConvLSTM) replaces the matrix products of the LSTM with convolutions, so its state is a spatial map;[18] this makes it a natural candidate for gridded storage fields.

The Nash-Sutcliffe efficiency (NSE) scales the squared error by the spread of the measured values,[19] which means that a model reproducing a strong seasonal cycle can score well without capturing anything else.[20] Skill scores based on mean squared error measure improvement over a chosen reference;[21] when a variable evolves slowly, the obvious yardsticks are its calendar-month average and simply carrying the previous value forward. Because consecutive months are dependent, random cross-validation inflates skill and chronological evaluation is preferable;[22] for the same reason, paired comparisons need resampling schemes that respect autocorrelation, such as the moving-block bootstrap.[23] Storage-based drought indices standardise the anomaly by its monthly statistics,[24,25] which allows categorical agreement to be tested in addition to continuous error.

The question we address is narrower than most of this prediction work and, we argue, more basic. A network that receives recent storage values can score well simply by repeating them, because storage changes slowly. Without a comparison against the naive rule of carrying the current month forward, nobody can tell what the network itself adds. The same caution applies to every preprocessing step that could carry information from the evaluation period into training.[26] Our aim is therefore to quantify the next-month storage signal that a convolutional recurrent model can draw from climate inputs alone over Tamil Nadu, with leakage controlled, and to set it against the information already held in the latest GRACE field. Our contributions are as follows:

we derive every GRACE month from the observation interval written into the band metadata and show that a count-based labelling with a manually entered gap list would assign a wrong month to 81 of the 238 fields;

we build a pipeline in which the study mask, scaling statistics, climatologies and drought references are computed from training months only, and confirm this with 14 automated checks run on the real data;

we train the same ConvLSTM with and without the latest storage field and compare both versions against persistence, climatology and a Random Forest using skill scores and block-bootstrap intervals over 19 test months;

we report per-mascon, regional and drought-class results alongside cell-level metrics, so that the small number of independent spatial units remains visible to the reader.

Several outcomes contradicted the impression given by headline accuracy, and they are reported as obtained. In particular, the variant with the lowest RMSE adds no statistically significant skill over persistence, although a test on 19 months has little power. Persistence and Model A both use the previous month's field, which would not yet be available in real time, so our experiments describe an idealised, zero-latency setting that we discuss explicitly. Figure 1 summarises the workflow.

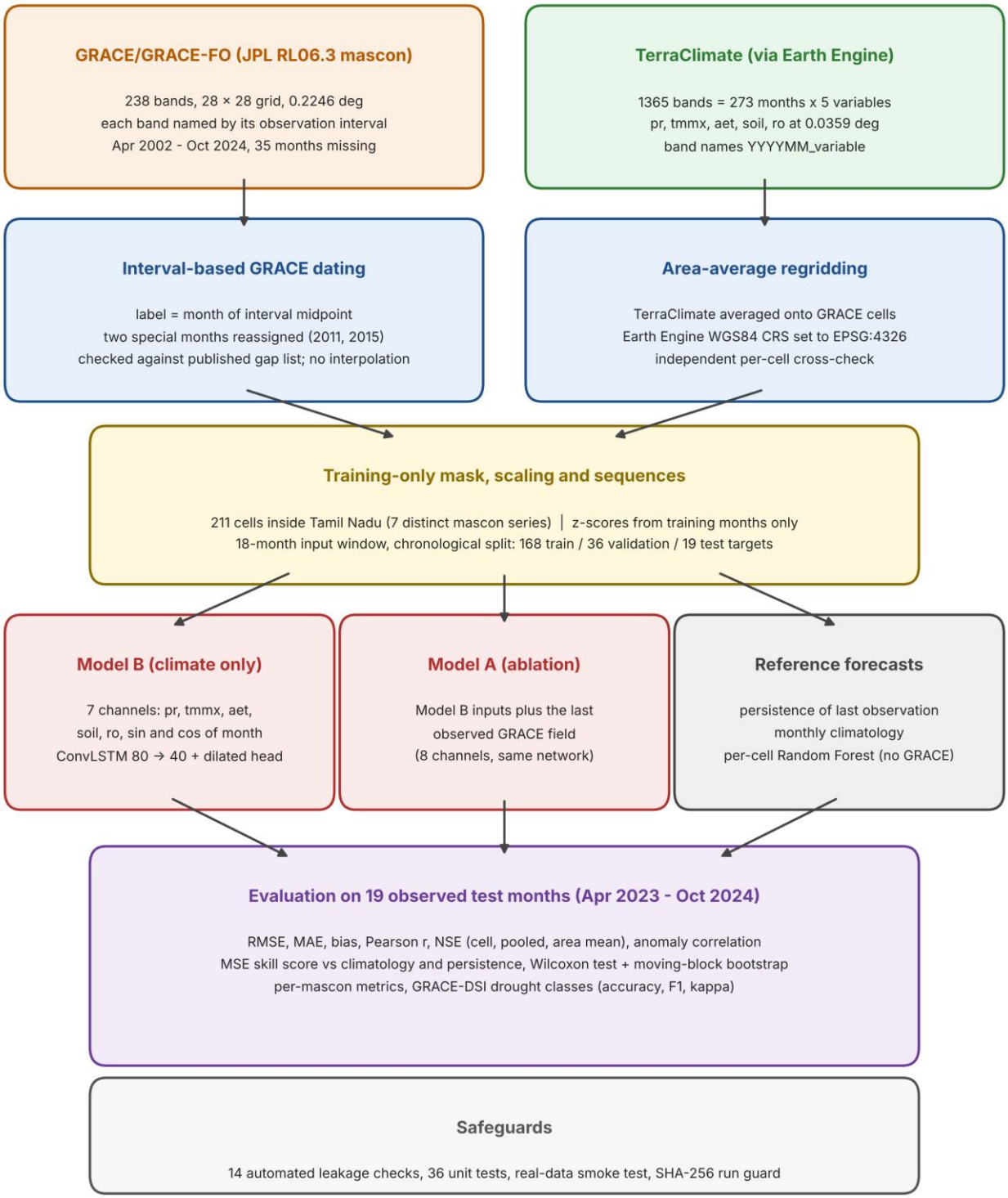


Figure 1. Workflow of the study. GRACE and TerraClimate exports are dated and regridded by explicit rules, a mask and scaling statistics are derived from the training period only, and two ConvLSTM variants are compared with three reference forecasts on 19 observed test months. Automated checks run before any model is trained.

Results

All scores refer to the 19 observed test months (April 2023 to October 2024) and the 211 cells of the study mask. We report the root-mean-square error (RMSE), the mean absolute error (MAE), the mean bias, the Pearson correlation (r), the NSE and the mean squared error skill score (MSESS); their definitions are given in Methods.

Data integrity and training

All data checks and leakage checks passed. The 238 GRACE labels were unique and increasing, the derived gaps matched the published list, and the two special months were reassigned as intended. No climate month was absent or incomplete, and the expected annual rhythm of rainfall and temperature verified that the variables had been read in the right order. Training losses fell steadily for all six runs, whereas validation losses reached their minimum early and then fluctuated, as expected with 168 training samples (Figure 2). Checkpoints were taken from epochs 26, 11 and 15 (Model B) and 29, 18 and 18 (Model A).

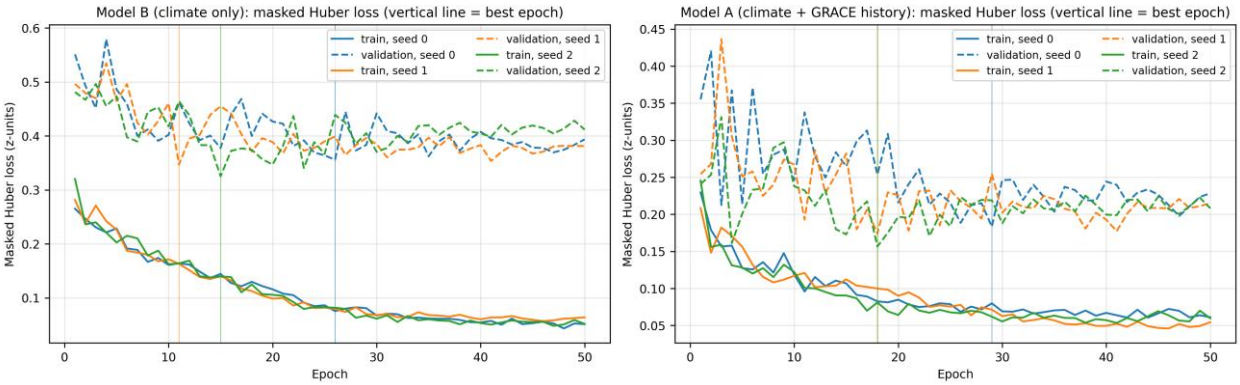


Figure 2. Learning curves (masked Huber loss, z-units) of each seed, climate-only network on the left and storage-assisted network on the right; training solid, validation dashed. Vertical lines mark the selected epoch of each seed.

Accuracy against the reference forecasts

Table 1 lists the test scores of every method and Figure 3 compares them graphically. Averaged over seeds, the climate-driven network missed by 7.84 cm in root-mean-square terms (0.645 z-units) and 6.49 cm on average, with r = 0.639 across all cell-months. Its area-mean series correlated with the observed one at 0.822, with an efficiency of 0.455. Cell-level efficiencies were mostly slightly negative: half the cells lay below −0.160 and the all-cell value was −0.139; a handful of border cells whose test-period values barely vary dragged the arithmetic mean to −2.092.

Table 1. Test scores on 19 observed months (April 2023-October 2024) and 211 cells. RMSE, MAE and bias in cm; NSE pooled over cell-months.

Method	RMSE	MAE	Bias	r	Median cell r	NSE	Area r	Area NSE
Model B (climate)	7.84	6.49	−2.47	0.639	0.785	−0.139	0.822	0.455
Persistence	5.12	4.13	+0.04	0.849	0.758	0.513	0.740	0.477
Climatology	12.76	10.75	−10.08	0.628	0.850	−2.020	0.867	−1.709
Random Forest	12.26	9.97	−8.43	0.527	0.761	−1.787	0.836	−1.052
Model A (+storage)	4.52	3.59	+0.60	0.880	0.891	0.621	0.896	0.772

Climatology and the Random Forest were far less accurate, with RMSEs of 12.76 and 12.26 cm and large negative biases (−10.08 and −8.43 cm). The test period was considerably wetter than the training average, and both methods are anchored to it: climatology by construction, and the

Random Forest because it learned mainly the seasonal cycle, its most important feature being the sine of the target month (importance 0.385). A forest also cannot predict outside the range of the targets it was trained on, which handicaps it in a wet test period, and its hyperparameters were not tuned. The network too fell short on average (−2.47 cm), yet its 18-month climate history allowed it to capture most of the shift towards wetter conditions.

Persistence was the strongest reference, with an RMSE of 5.12 cm, negligible bias and a pooled correlation of 0.849. Model B exceeded persistence on two correlation measures, the state-average series (0.822 versus 0.740) and the median per-cell correlation (0.785 versus 0.758), a consequence of persistence trailing every seasonal reversal by one month. Climatology scored higher still on both (0.867 and 0.850), so these measures mainly reflect the seasonal cycle and not forecast skill. Model A achieved the best score in Table 1 on every measure except bias, for which persistence was closer to zero: its errors fell to 4.52 cm (RMSE), its correlation rose to 0.880, and the state-average efficiency reached 0.772. Scatter plots of every method are collected in Figure 4.

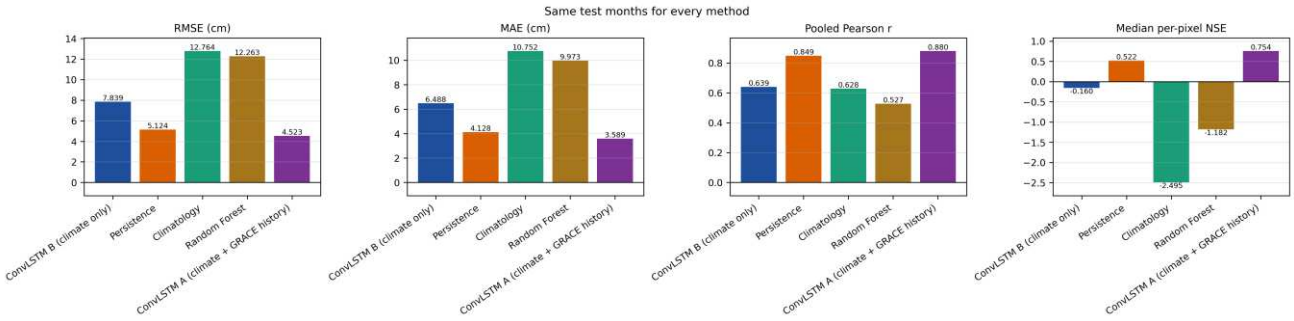


Figure 3. Error and agreement scores (RMSE, MAE, r, median cell NSE) for each method, all computed on the same months and cells.

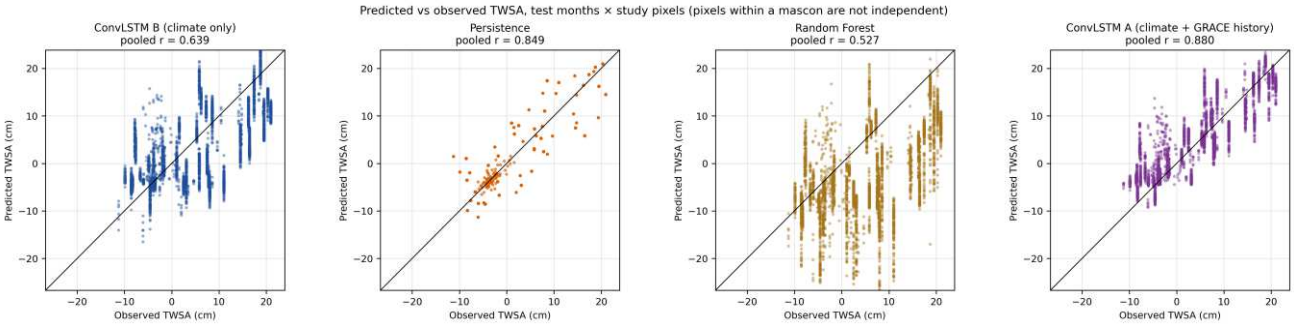


Figure 4. Scatter of forecasts versus GRACE values over the test cell-months, shown in turn for the climate-only network, the persistence rule, the Random Forest and the storage-assisted network. Points from cells of one mascon are not independent.

Seed variability

The three Model B seeds differed moderately (Table 2). Individual runs gave RMSEs between 7.68 and 8.73 cm and correlations between 0.579 and 0.650; the run initialised with seed 2 lagged behind. Averaging the seeds improved on their mean RMSE (7.84 against 8.13 cm), which is why the ensemble is reported while each seed remains visible.

Table 2. Model B results per seed on the test months.

Seed	RMSE (cm)	MAE (cm)	Bias (cm)	r	Median NSE	Area r
0	7.96	6.54	−3.35	0.647	−0.088	0.818
1	7.68	6.26	−2.29	0.650	0.023	0.847
2	8.73	7.26	−1.78	0.579	−0.463	0.775
Mean ± SD	8.13 ± 0.54	6.69 ± 0.52	−2.47 ± 0.80	0.626 ± 0.040	−0.176 ± 0.254	0.813 ± 0.036
Ensemble	7.84	6.49	−2.47	0.639	−0.160	0.822

Behaviour through the seasonal cycle

Measured state-average storage climbed from roughly zero in mid-2023 to some 14 cm by December, once the winter monsoon had passed, dropped near −9 cm by April 2024 and rebounded to around 10 cm in October (Figure 5). Model B placed the turning points correctly but misjudged the level: it ran 5-10 cm low from April to August 2023 and overshot the December peak, predicting about 19 cm. Persistence followed the curve closely but one month late. Climatology and the Random Forest reproduced the shape of the cycle while staying well below the observations, and Model A tracked the observations most closely. Persistence and Model B therefore err in different months: the former at transitions, the latter in the level of storage.

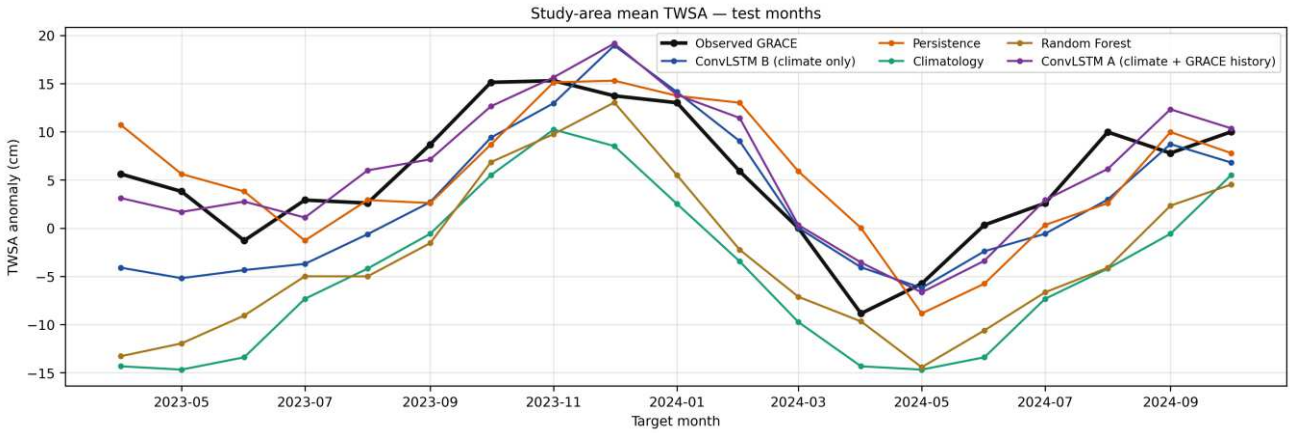


Figure 5. Area-mean storage anomaly (cm) over the 19 observed test months for the observations, both ConvLSTM variants and the three references.

Significance and skill

Table 3 gives the paired monthly comparisons. In 16 of 19 months the climate-only network beat the calendar-mean forecast, and the resampled interval of the difference stayed entirely below zero; the same held against the Random Forest (Wilcoxon $p = 3.8 \times 10^{-5}$; interval −0.347 to −0.211; mean absolute errors of 0.534 and 0.821). Persistence, by contrast, was outdone in just 4 months. For this comparison the evidence is mixed: the Wilcoxon test indicated that Model B was less accurate ($p = 0.011$), but once autocorrelation was respected the bootstrap interval, −0.022 to 0.300, included zero. By our own criterion that a difference is robust only if the 95% interval excludes zero, Model B being worse than persistence is therefore suggestive and not established. The storage-assisted network won 12 of the 19 monthly contests with persistence, and neither test found a difference ($p = 0.374$; interval −0.152 to 0.034). With 19 autocorrelated months these tests have little power, so the result means that no advantage was

detected, not that the two forecasts are equivalent. The paired tests use standardised monthly area-mean errors, whereas the headline scores are cell-level errors in centimetres, and the two criteria need not agree.

Table 3. Paired comparisons of monthly area-mean MAE (standardised units, n = 19 months) with 95% moving-block bootstrap intervals of the mean difference (block 3, 10,000 resamples).

Comparison	MAE 1	MAE 2	W	p	95% interval	Months 1 better
B vs persistence	0.534	0.340	33	0.0108	[−0.022, 0.300]	4/19
B vs climatology	0.534	0.885	7	7.2 × 10−5	[−0.480, −0.214]	16/19
A vs persistence	0.295	0.340	72	0.374	[−0.152, 0.034]	12/19

A negative difference favours the first method. W is the Wilcoxon statistic. Mean absolute errors (MAE) are monthly area means in standardised units, not centimetres.

Table 4 gives the skill scores together with a split of the mean squared error into its squared bias and the remaining error variance. Model B removed 62% of the squared error of climatology (MSESS 0.623, interval 0.487-0.729) but scored −1.340 against persistence (interval −2.577 to 0.018, which just includes zero). The split shows that most of the gain over climatology is a bias correction: the squared bias of climatology is about 102 of its 163 cm^2 mean squared error, against about 6 of 61 cm^2 for Model B, so roughly 94% of the reduction in mean squared error comes from recognising that 2023-2024 was wetter than the training period. With bias removed, the remaining error falls from about 61 to 55 cm^2, an improvement of roughly 10%. Subtracting the training-period monthly means from forecasts and observations leaves anomaly correlation coefficients (ACC) of 0.365 (all cell-months), 0.369 (median cell) and 0.381 (state mean); the bulk of the raw agreement therefore stems from the annual cycle and the elevated wet-period baseline. The ACC was evaluated for Model B only.

Table 4. Mean squared error skill score of each method relative to climatology and to persistence, with the mean squared error split into squared bias and remaining error variance.

Method	MSESS vs climatology	MSESS vs persistence	MSE (cm^2)	Squared bias (cm^2)	Remaining (cm^2)
Model B	0.623 [0.487, 0.729]	−1.340 [−2.577, 0.018]	61.5	6.1	55.4
Persistence	0.839	-	26.2	0.0	26.2
Random Forest	0.077	−4.728	150.3	71.1	79.2
Model A	0.874	0.221	20.4	0.4	20.1
Climatology	-	-	162.8	101.6	61.2

Bracketed values are 95% moving-block bootstrap intervals, computed for Model B only. Mean squared error (MSE) is pooled over all test cell-months; the squared bias is the square of the mean error and the remainder is the variance of the errors. MSE and squared bias were derived from the RMSE and bias in Table 1 and are rounded.

Contribution of the storage channel

Supplying the latest observed storage field (Table 5, Figure 6) reduced the RMSE from 7.84 to 4.52 cm and raised the median cell efficiency from −0.160 to 0.754. Judged against Model B, the change is large. Judged against persistence, which reaches 5.12 cm without any training, most of it is explained by the memory carried in the extra channel; the positive skill score of Model A over persistence (0.221) is not statistically significant, although with 19 months a modest gain could not have been detected. Because persistence and Model A both use the previous month's field, which would arrive with a delay in practice, these results describe an idealised upper bound on the benefit of storage memory.

Table 5. Ablation of the storage channel.

Metric	Model B	Model A	Persistence
Input channels	7	8	-
RMSE (cm)	7.84	4.52	5.12
MAE (cm)	6.49	3.59	4.13
Pooled r	0.639	0.880	0.849
Median cell NSE	−0.160	0.754	0.522
Area-mean r	0.822	0.896	0.740
Area-mean NSE	0.455	0.772	0.477
Better than persistence (robust)?	No	No	-

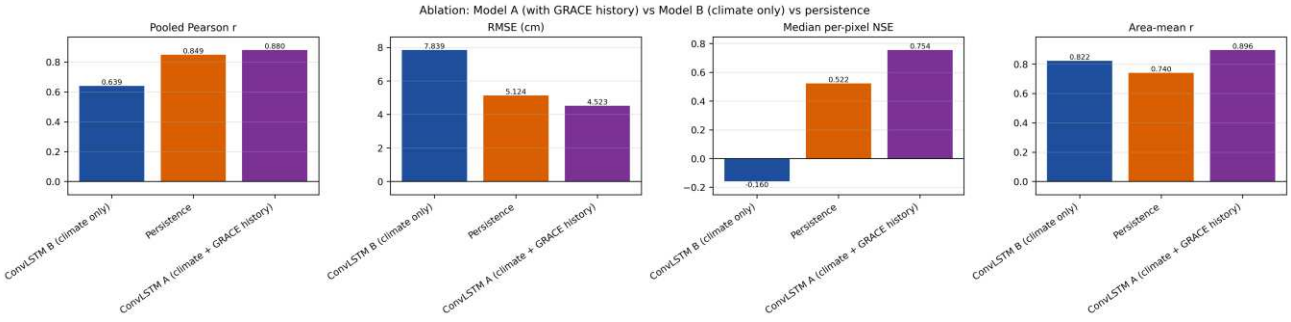


Figure 6. Model A (climate plus latest storage), Model B (climate only) and persistence compared on pooled correlation, RMSE, median cell NSE and area-mean correlation.

Spatial structure

Observed fields show a sharp change near 10.5 degrees north, where the northern and southern mascons meet and often carry anomalies of opposite sign (Figure 7). Model B produced smooth fields that matched the overall level but blurred this boundary. April 2023 came out uniformly drier than measured, whereas October 2024 showed an underestimate in the north alongside an overestimate in the south. Because the annual cycle is shared statewide, temporal correlation stayed high in nearly every cell (median 0.785); the largest absolute errors, by contrast, concentrated in the north-central block of the biggest mascon, where storage had risen most (Figure 8).

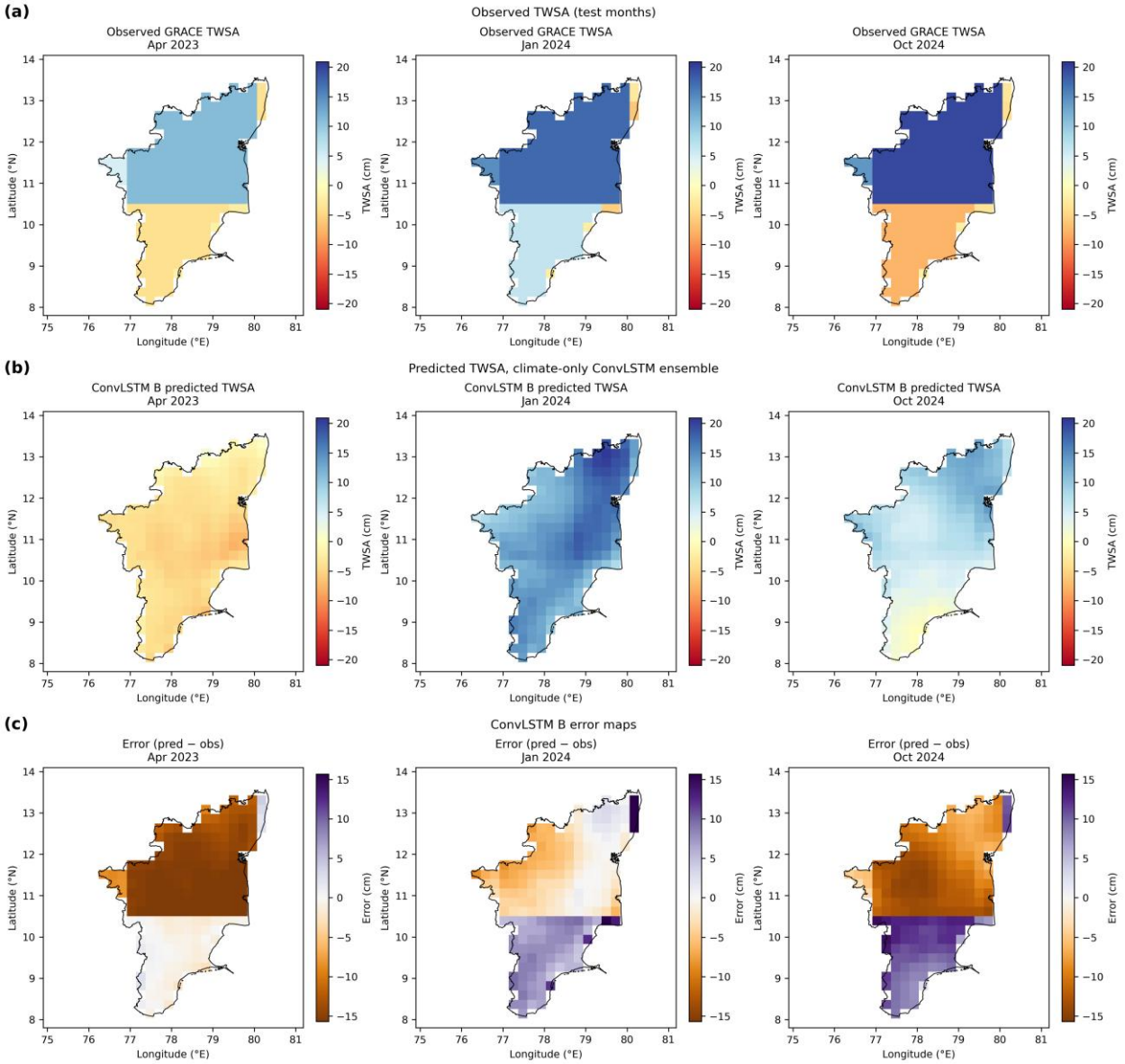


Figure 7. Observed storage (a), Model B prediction (b) and error (c), in cm, for April 2023, January 2024 and October 2024.

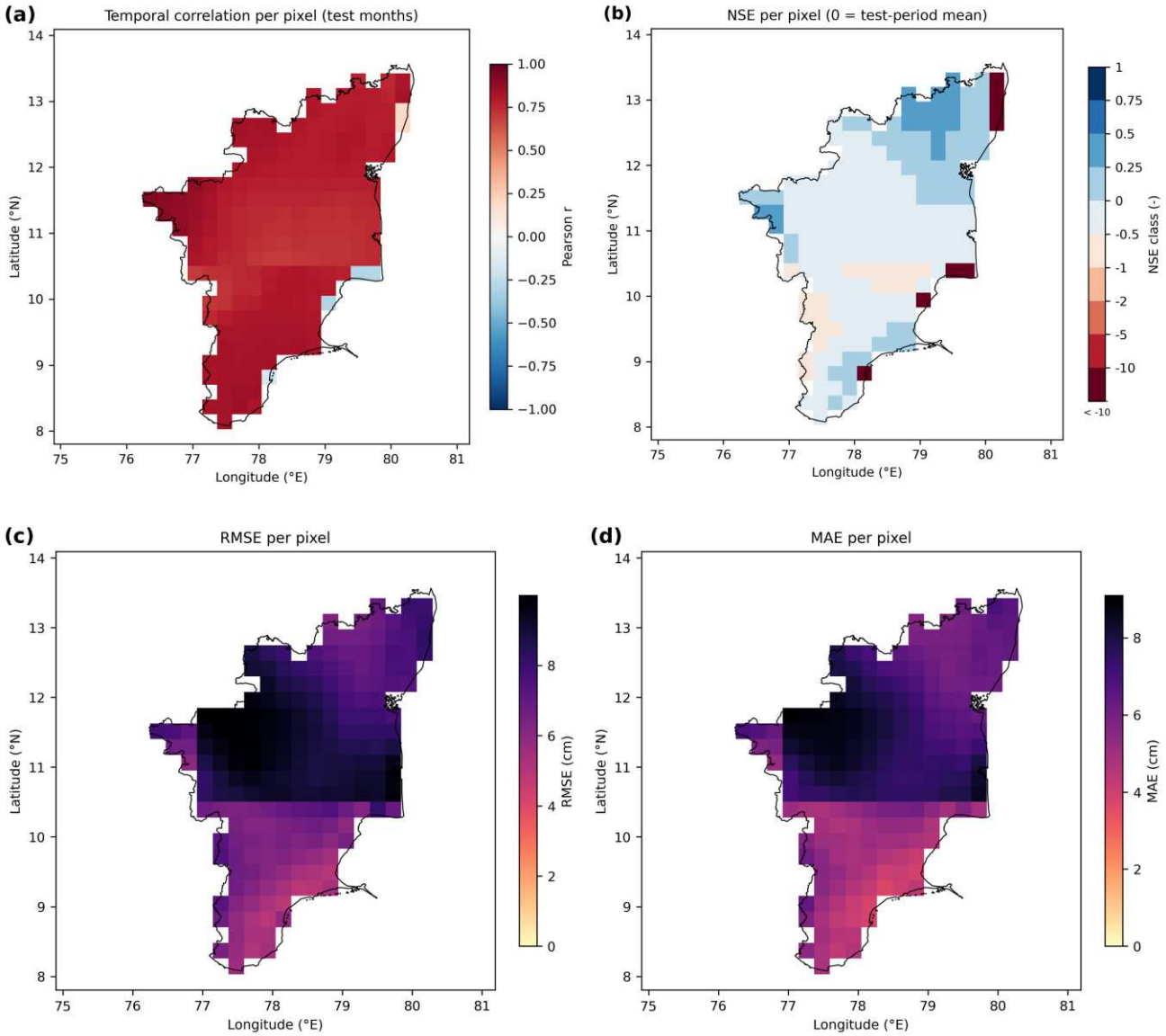


Figure 8. Cell-by-cell scores of the climate-only network: temporal correlation (a), NSE (b), RMSE in cm (c) and MAE in cm (d). In (b) the lowest colour class is open-ended (minimum −88.62).

Per-mascon scores are given in Table 6. Model B correlated at 0.784-0.901 with the three core mascons, with efficiencies between −0.287 and 0.256 and a cell-weighted mean of −0.094 (correlation 0.796). The four two-cell edge fragments scored very poorly because their test variance is small; they are listed for transparency. Dividing the grid into overlapping northern, southern, western and eastern portions (Table 7, Figure 9) showed Model B weakest in the north, where persistence and Model A remained far ahead. In the southern sector persistence (0.725) outperformed Model A (0.631).

Table 6. Per-mascon test scores of Model B.

Mascon	Cells in mask	NSE	r
M0 (core)	6	0.256	0.901
M1 (core)	131	−0.012	0.784
M2 (core)	66	−0.287	0.811
Fragments (four)	2 each	−31.7; −77.6; −19.2; −76.1	0.185; 0.817; −0.239; −0.288
Core, cell-weighted	203	−0.094	0.796

Table 7. Pooled efficiency (1 – SSE/SST over all cell-months of a sector) in four overlapping grid sectors, which are not basins.

Method	North (115 cells)	South (96)	West (81)	East (130)
Model B	0.040	0.266	0.282	0.307
Persistence	0.560	0.725	0.714	0.691
Model A	0.778	0.631	0.773	0.762

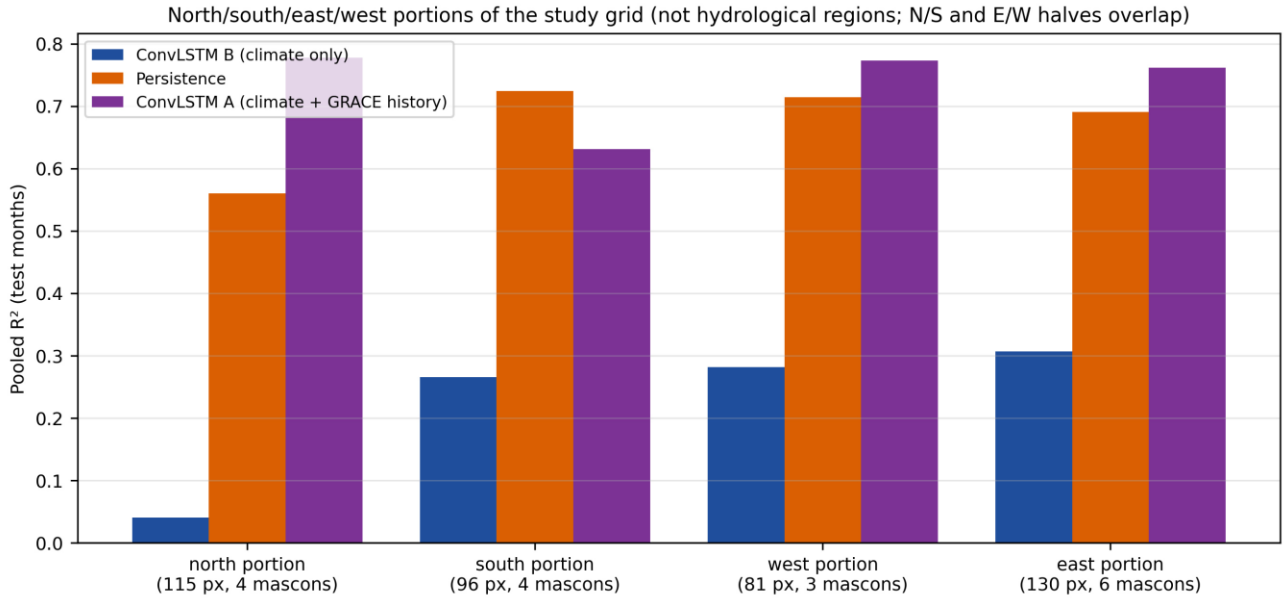


Figure 9. Pooled efficiency (1 – SSE/SST; labelled pooled R^2 in the plot) of each method in the four grid sectors.

Drought classes

Drought and wetness classes were derived from the GRACE drought severity index (GRACE-DSI; see Methods). Because 2023-2024 was humid by training-period standards, the upper two classes dominate the test set; dry conditions emerge only in places, chiefly an area spanning roughly a third of the state during September-October 2024. Model B reached an accuracy of 0.339 and a Cohen's kappa of 0.030, essentially chance agreement. Persistence reached an accuracy of 0.619 (kappa 0.428) and Model A 0.653 (kappa 0.445) (Table 8); persistence still had the higher macro F1 score (0.569 versus 0.393), and the weighted F1 scores were almost identical (0.626 versus 0.625). Its confusion matrix shows that very wet cell-months were mostly assigned to the wet or normal class (Figure 10), and the class maps and drought extent

show that it missed the 2024 patch, which persistence registered one month late (Figures 11 and 12).

Table 8. Agreement of GRACE-DSI drought and wetness classes over 4009 test cell-months.

Index	Method	Accuracy	F1 macro	F1 weighted	Kappa
GRACE-DSI	Model B	0.339	0.246	0.333	0.030
GRACE-DSI	Persistence	0.619	0.569	0.626	0.428
GRACE-DSI	Model A	0.653	0.393	0.625	0.445

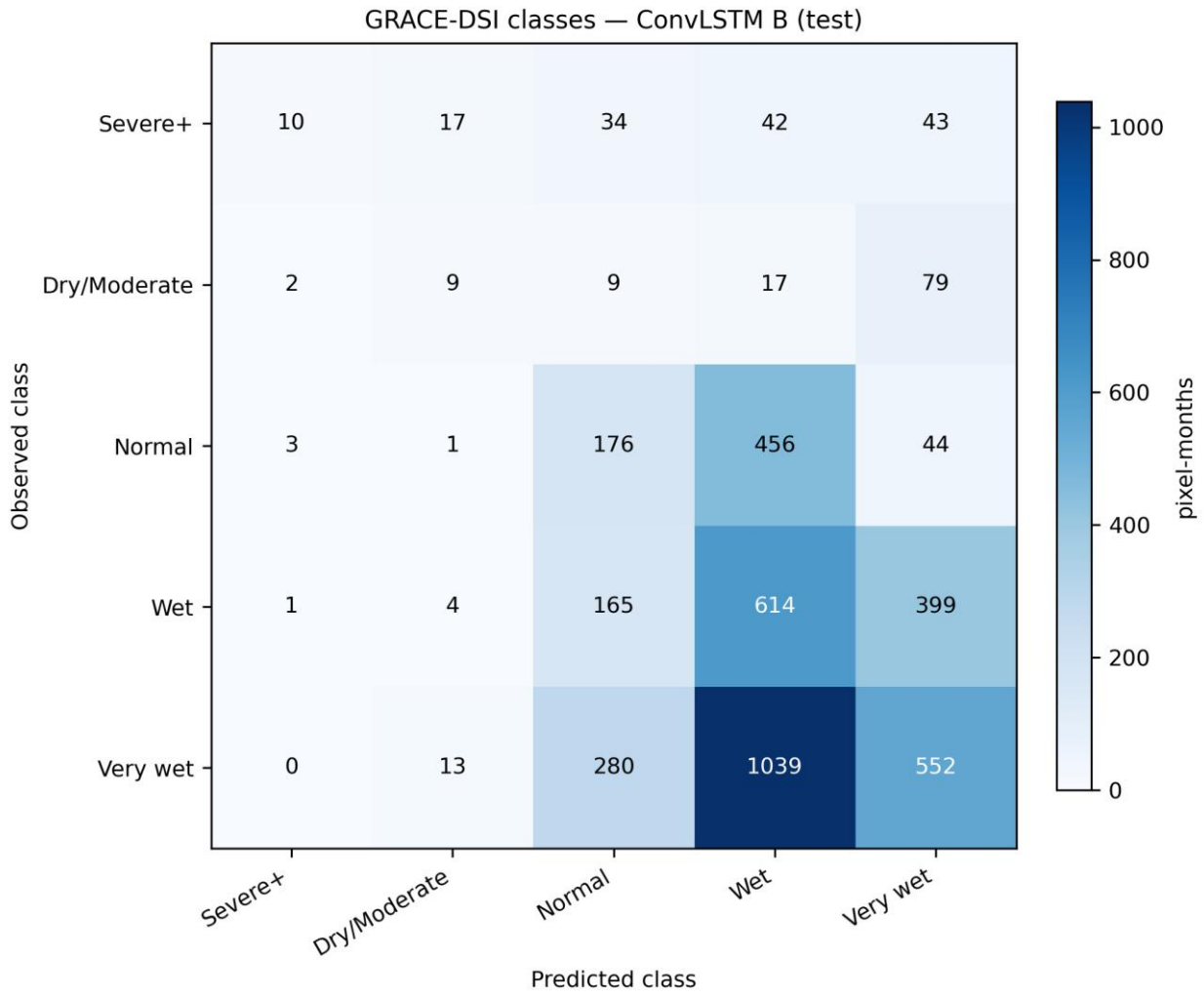


Figure 10. Confusion matrix of GRACE-DSI classes for Model B over the test cell-months (rows observed, columns predicted).

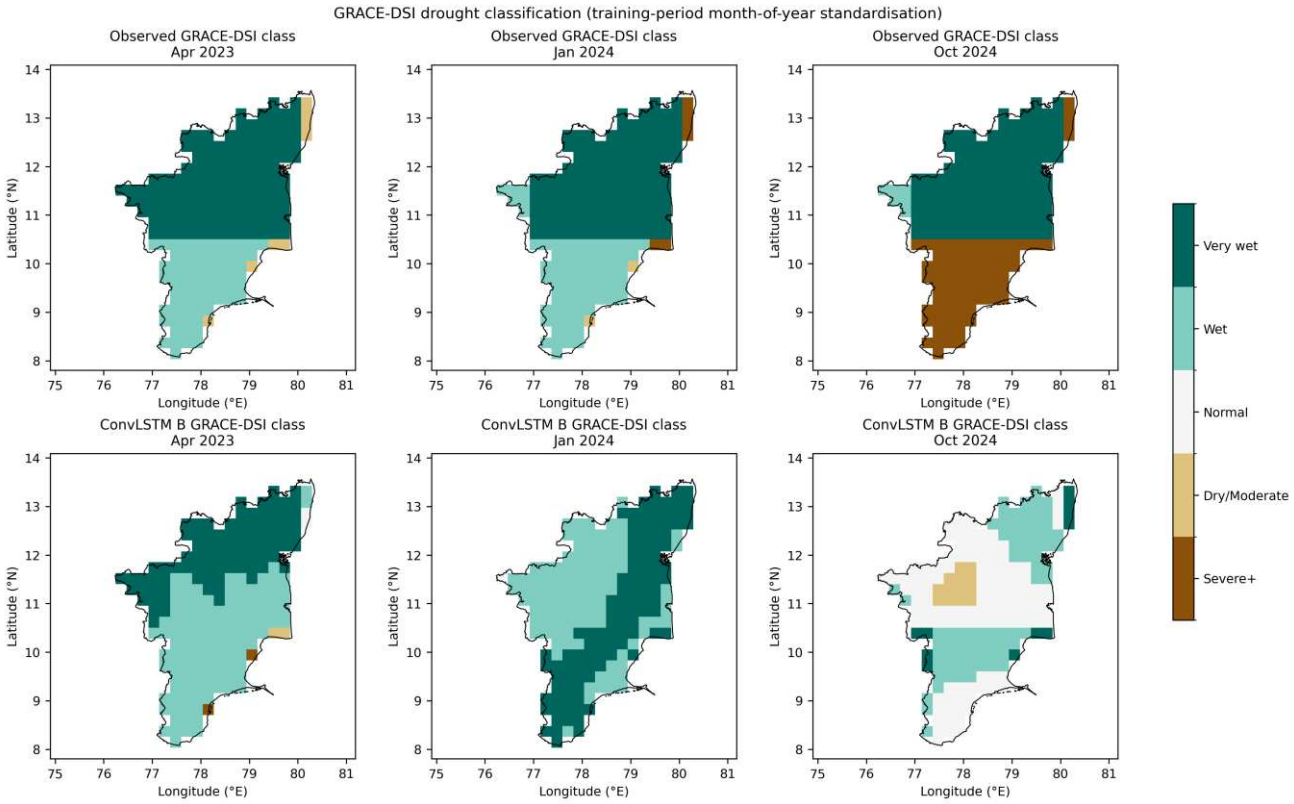


Figure 11. Drought-index categories derived from GRACE (upper row) and from the climate-only forecasts (lower row) in three sample months.

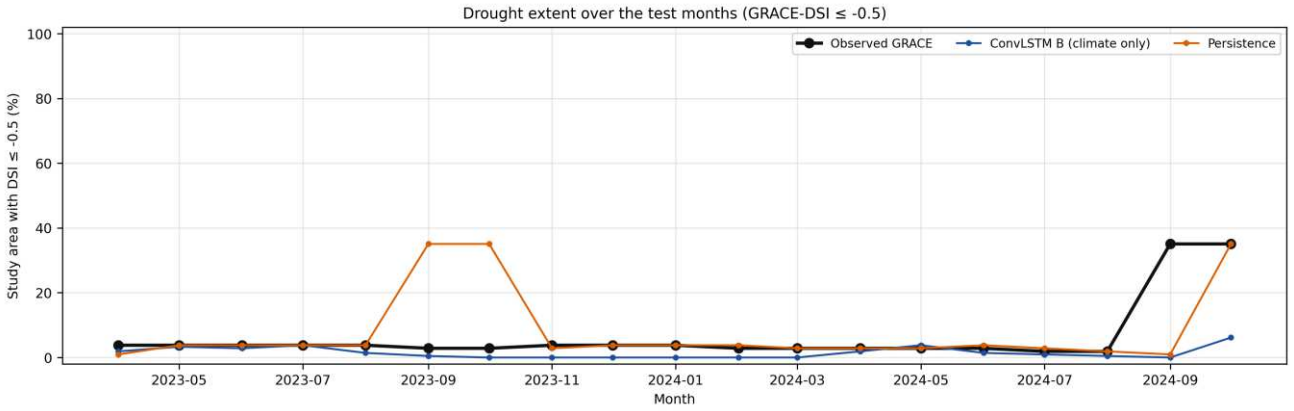


Figure 12. Fraction of the state classed as dry (index at or below −0.5) through the test period, observed and forecast.

Discussion

What the climate-only network learns. Model B outperforms monthly climatology and an untuned per-cell Random Forest in RMSE, so the 18-month climate history holds information about the level of storage that these references miss. Most of this advantage, however, is a bias correction: about 94% of the reduction in mean squared error relative to climatology comes from recognising that 2023-2024 was wetter than the training average, and with bias removed the remaining error falls by only about 10%. Its anomaly correlation of 0.365 and near-chance drought-class agreement (kappa 0.030) show that its grasp of month-to-month departures from the seasonal cycle is limited. We therefore read the climate signal as informative about the wet shift but weak for month-to-month anomalies. The network has about 430,000

parameters, was trained on 168 samples and is evaluated on three core mascons, and we did not test a version without spatial convolutions, so we cannot attribute its advantage over the Random Forest to its spatial structure.

Why persistence is hard to beat. Because storage accumulates the effect of many months of rain, consecutive months differ little compared with the annual swing, and copying the latest value fails mainly when the cycle reverses. Outperforming it requires anticipating the month-on-month increment, and that is exactly the part of the signal climate inputs explain least well. For this state and a lead time of one month, climate inputs by themselves fell short, although the evidence that Model B is less accurate than persistence rests on the Wilcoxon test alone, because the block-bootstrap interval includes zero. The errors of the two approaches fall in different months, which suggests that predicting the change from climate and adding it to the last observation could combine their strengths.

Reading the storage-assisted variant. Model A has the lowest RMSE and the highest pooled correlation in Table 1, but it is not best on every measure: persistence is ahead of it in the southern sector (Table 7) and in macro F1 for the drought classes (Table 8). A comparison limited to Model B or to climatology would credit the network with strong forecasting skill. The persistence comparison removes that impression: the extra channel supplies the information and the network adds no statistically detectable skill beyond it. With 19 autocorrelated months from a single wet period this test has low power, so the result is an absence of evidence for an advantage and not evidence that the two forecasts are equivalent. Published accuracies of models that receive lagged storage, or quantities derived from it, should be read with this in mind and, where possible, compared with persistence.

Practical value. Gravity fields reach users late, some months are absent altogether, and another break between missions cannot be ruled out. In these circumstances an estimate that removes most of the error of climatology may be useful, although most of that gain is a correction of bias in a wet period. Persistence and Model A use the previous month's field, which would not be available in real time. The practical setting, in which the newest field is several months old and forecasts are needed several months ahead, is where climate inputs may contribute most; we did not test it (see Limitations). If a fresh field is already in hand, nothing in our results argues for replacing persistence with the network at a one-month lead.

The weight of preprocessing. The dating of the storage fields shows how preprocessing can outweigh modelling choices. A count-based labelling would have shifted 81 of 238 fields, every GRACE-FO band by two months, an error that is invisible in summary metrics but alters every downstream number. Reading dates from metadata, cross-checking the regridding and automating leakage tests should therefore be treated as part of the method rather than as housekeeping.

Limitations

The target is total storage; groundwater was not separated and no well observations were used, so the results say nothing direct about aquifers. The evaluation rests on 19 autocorrelated months from one unusually wet period and on seven distinct spatial series, three of them core, which keeps intervals wide and leaves performance during a dry spell untested. A rolling-origin

evaluation that includes dry years, for example the 2016-2017 north-east monsoon failure that lies in our training period, is the most important next step. Only a one-month horizon and a single architecture were examined. Persistence and Model A use the previous month's field, so the experiments assume zero latency; forecasts from a field that is several months old, at leads of one to six months, were not run.

The set of reference forecasts is also limited. We did not test persistence combined with the climatological seasonal change, a linear model on lagged climate or an LSTM on the state-mean series, and the Random Forest was not tuned and cannot extrapolate beyond its training targets. Reporting is incomplete in places: the paired tests use standardised monthly area-mean errors, the skill-score intervals were computed for Model B only, and the anomaly correlations are reported for Model B only. The Earth Engine collection we used is the mascon grid without the coastal-resolution-improvement (CRI) filter, which does not correct leakage across coastlines;[27] this may affect coastal cells in a state with a long coastline, and the analysis should be repeated with the CRI product, which is recommended for non-expert users, and with GRACE-FO months released after the end of that collection. The special-month rule follows gravity-toolkit and could not be cross-checked against the official table of the Physical Oceanography Distributed Active Archive Center (PO.DAAC) in our environment, and the sensitivity of the results to the drought-class thresholds was not examined.

Finally, the scope of the evidence is narrow. One state, one lead time, one architecture and three random seeds were examined, so the numbers reported here should not be read as a general statement about deep learning for storage prediction, nor as transferable to other regions, resolutions or lead times. Because all cells within a three-degree mascon cap share one value, the 211 cells of the mask carry only seven distinct series, three of them core; the effective sample size behind every score is therefore far smaller than the cell count suggests, and this is the main reason the confidence intervals are wide and the paired tests have little power. The ConvLSTM hyperparameters were fixed a priori rather than searched, the Random Forest was left at its defaults, and no formal uncertainty estimate accompanies the predictions, so the comparison between methods reflects these particular configurations rather than the best attainable version of each. We also report no measure of how the networks would behave if a GRACE month were missing at prediction time, a situation the record makes likely.

Conclusions

Once dates, regridding and data splits are verified and every statistic is confined to the training period, a ConvLSTM driven only by TerraClimate forcing predicts next-month GRACE storage over Tamil Nadu with a lower RMSE (7.84 cm) than climatology (12.76 cm) or a per-cell Random Forest (12.26 cm), mainly by correcting the bias of a wet period, but its RMSE is higher than that of persistence, a difference supported by the Wilcoxon test and not by the bootstrap interval. Adding the latest storage field gives the lowest RMSE, yet its gain over persistence is not statistically significant in 19 test months. Month-ahead skill reported for storage models should therefore always be measured against persistence. Further work will examine longer horizons, prediction of the monthly change, additional predictors such as reservoir storage and climate indices, rolling-origin evaluation that includes dry years, latency-aware experiments with leads of one to six months, additional baselines, the CRI product, newer GRACE-FO months, and

validation against well observations. These conclusions are drawn from a single state, a single lead time and a single architecture, and we make no claim beyond that setting.

Methods

Figure 1 outlines the workflow. Both inputs are exported from Google Earth Engine, dated and aligned by explicit rules, restricted to a study mask defined without evaluation-period information, and converted into 18-month sequences. Two ConvLSTM variants and three reference forecasts are then compared on the same test months.

Study region

Tamil Nadu occupies roughly 130,000 km^2 of south-eastern peninsular India between about 8.1 and 13.6 degrees north. The western interior collects a share of its rain in the June-September (south-western) monsoon, whereas the eastern and coastal districts rely chiefly on the October-December (north-eastern) rains. Hard-rock aquifers of limited capacity underlie much of the state, and irrigation tanks and groundwater supply a large share of agricultural and urban demand. Storage typically bottoms out before the north-east monsoon and peaks around December or January. The state outline was taken from the DataMeet administrative boundaries[28] (CC BY 4.0) and simplified with a tolerance of 0.005 degrees.

Storage data

Our target variable is the equivalent water thickness supplied by version 4 of the Jet Propulsion Laboratory (JPL) RL06.3 mascon product for both missions.[29] We obtained it through Earth Engine[27,30] as the collection NASA/GRACE/MASS_GRIDS_V04/MASCON, which is the mascon grid without the CRI filter and does not correct leakage across coastlines. Each value is a departure from the 2004-2009 mean. The exported GeoTIFF holds 238 bands on a 28 × 28 grid at 0.2246 degrees in EPSG:4326, and each band description states the first and last day of its solution interval, for example 20020403_20020429. Thirty-five of the 273 months between April 2002 and December 2024 lack a field, comprising the known isolated outages, the break between missions and the final two months of 2024, which are absent from the Earth Engine collection (newer GRACE-FO months may be available from PO.DAAC but were not used). Figure 13 shows the mascon structure and Figure 14 the months available in each period.

JPL mascon solutions estimate mass changes directly on equal-area spherical caps of about three degrees, with geophysical constraints applied during the inversion.[9] When the solution is exported on a regular grid finer than the caps, neighbouring cells repeat one series. Over our domain the 784-cell export holds only 14 distinct series, seven of which intersect the state. Any statistic computed per cell must be read with this redundancy in mind.

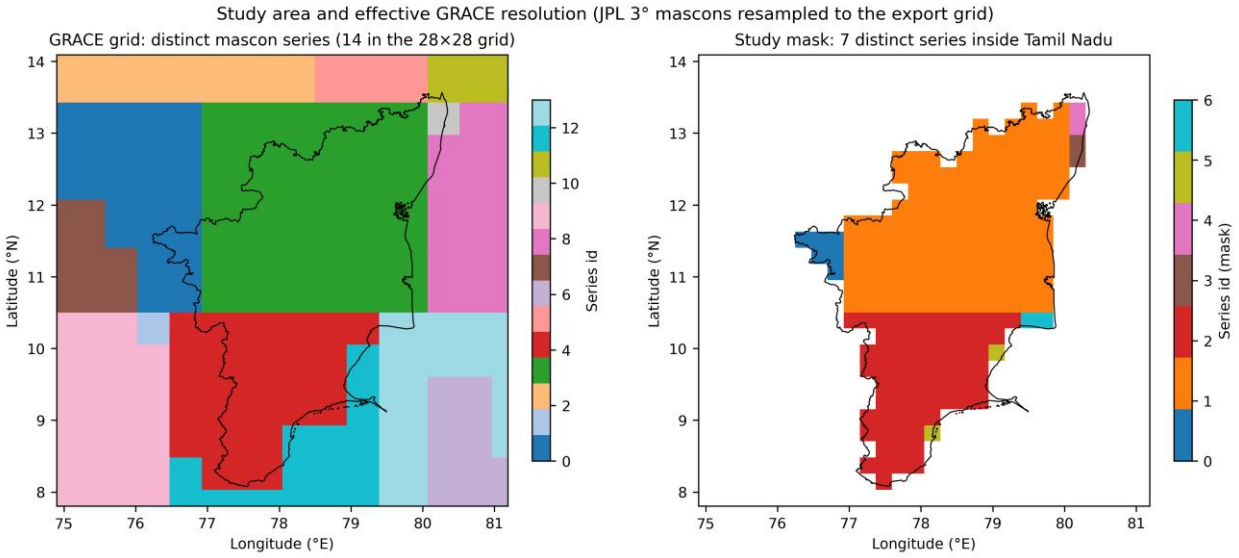


Figure 13. Unique storage series across the exported 28 × 28 grid (left); seven of them intersect the study mask (right). Two mascons cover most of the state; four series contribute only two cells each.

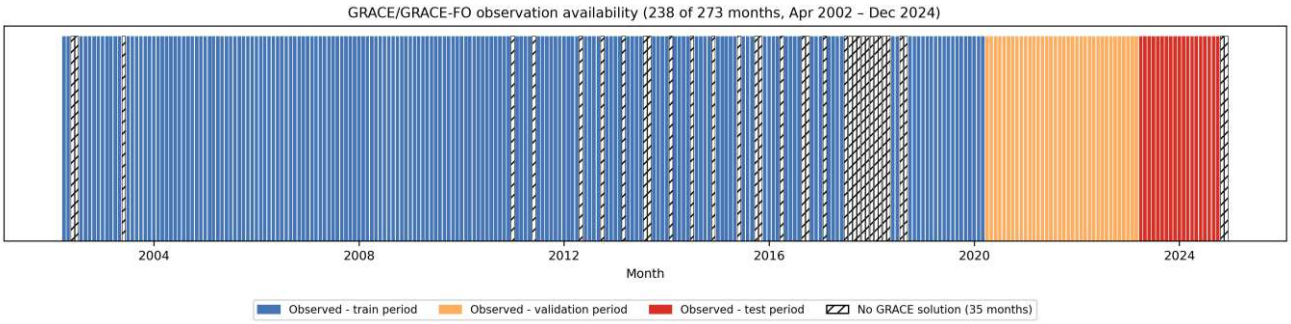


Figure 14. Months with a GRACE/GRACE-FO solution between April 2002 and December 2024 (238 of 273), coloured by the period to which the target month belongs. Hatched months have no solution and are never used as targets.

Climate forcing

TerraClimate[31] combines climate normals with reanalysis anomalies to give monthly fields, including simple water-budget terms, at a native resolution of 1/24 degree (about 4.6 km). Five of them were used: precipitation (pr), mean maximum temperature (tmmx), actual evapotranspiration (aet), end-of-month soil moisture (soil) and runoff (ro). Together they describe input, demand, loss and part of the storage itself. The export contains 1365 int16 bands (273 months x 5 variables) on a 169 × 168 grid at 0.0359 degrees (about 4 km), named YYYYMM_variable. Earth Engine resampled the native 1/24-degree fields to this grid during export. Earth Engine labels its reference frame only as a nameless latitude-longitude system on WGS84; this exact case is recognised and assigned EPSG:4326, while any other system would be left unchanged. Sea pixels carry zeros instead of a missing-value flag, so any pixel that stays constant for the whole record was treated as invalid. Table 9 summarises both inputs.

Table 9. Properties of the two exports as recorded by the pipeline.

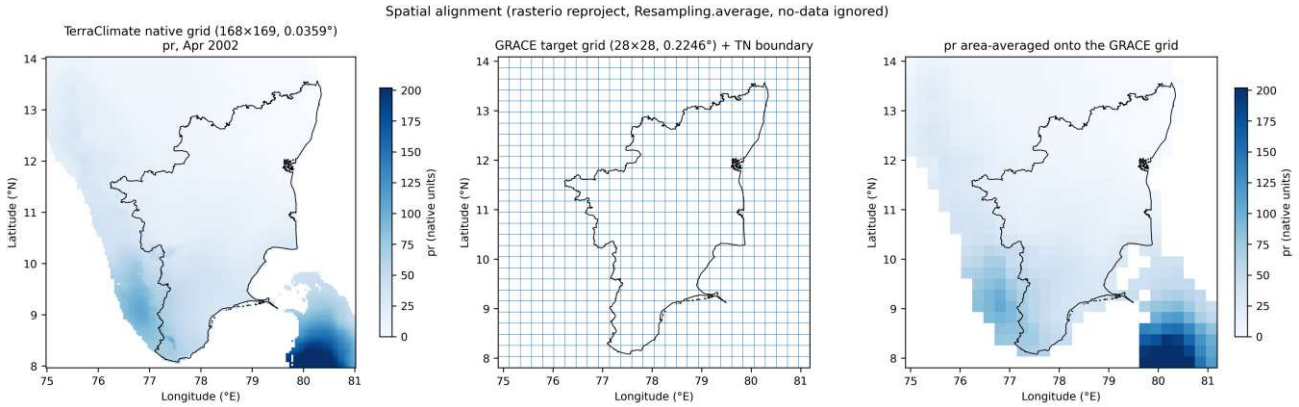
Property	GRACE/GRACE-FO	TerraClimate
Product	JPL RL06.3 mascon v4 (no CRI filter)	TerraClimate monthly
Variables	lwe_thickness (cm)	pr, tmmx, aet, soil, ro
Bands	238	1365
Grid	28 × 28	169 × 168
Cell size	0.2246 deg	0.0359 deg (resampled from 1/24 deg on export)
Coordinate system	EPSG:4326	WGS84 geographic (set to EPSG:4326)
Band naming	observation interval	YYYYMM_variable
Period	Apr 2002-Oct 2024 (observed)	Apr 2002-Dec 2024
Missing months	35 (not interpolated)	none

Dating the storage fields

For each band the interval start and end were parsed and the band was assigned to the calendar month containing the interval midpoint. Two solutions flown during reduced battery operation share a midpoint month with their neighbours and were reassigned following the rule implemented in the gravity-toolkit library:[32] 16 October-15 November 2011 to November 2011 and 11 April-10 May 2015 to May 2015. The pipeline then verifies that labels are unique and strictly increasing, that each label overlaps its own interval, and that the gaps up to 2018 coincide with the published list. Under this procedure the record ends in October 2024. To show why this step matters, we also tried a count-based alternative in which the n-th band is assigned to the n-th month of a manually entered list of gap months; it gave 81 of the 238 bands a wrong month and displaced every GRACE-FO band by two months. Missing months were never interpolated: months lacking a field are simply skipped as targets, and both persistence and Model A draw only on genuinely measured earlier fields.

Regridding and study mask

Each TerraClimate field was area-averaged onto the GRACE grid with rasterio's averaging resampler, ignoring invalid cells; about 39 native cells contribute to each GRACE cell. The covered fraction of every GRACE cell was computed simultaneously, and the averaged values were compared with an independent calculation over native cells whose centres fall inside each target cell (Figure 15). A cell entered the study mask only if its centre lay inside the state, at least half of its area was covered by valid TerraClimate data, all of its climate values were finite, and storage values existed in all measured training months. That final rule looks only at training data, never at later years. The mask holds 211 cells; one coastal cell was rejected for insufficient coverage.



441

442 **Figure 15. Regridding check for one month: precipitation at its original 4 km spacing (left),**
443 **the coarser storage grid with the state boundary (centre) and the same rainfall after aver-**
444 **aging onto it (right).**

445 Sequences, scaling and splits

446 Let $\mathbf{X}_j$ denote the C × H × W stack of input channels for month j and $\mathbf{Y}_t$ the observed storage
447 field for target month t. A model f maps the preceding 18 months to a prediction $\mathbf{P}_t$:

448 $$\mathbf{P}_t = f(\mathbf{X}_{t-18}, \dots, \mathbf{X}_{t-1}). \quad (1)$$

449 Calendar position enters through two channels computed from the month number m_j:

450 $$u_j = \sin(30^\circ \times m_j), \qquad v_j = \cos(30^\circ \times m_j). \quad (2)$$

451 Model B uses seven channels (pr, tmmx, aet, soil, ro, u, v). Model A adds an eighth containing,
452 for every input month, the latest observed storage field at or before that month, so that when
453 predicting month t it sees the value of month t - 1 or the most recent earlier solution. Each cli-
454 mate input was converted to z-scores using moments taken from masked cells in training
455 months alone; storage was standardised with statistics from observed training targets. Off-
456 mask cells were set to zero and given zero weight everywhere. Samples were divided by target
457 month into contiguous periods (Table 10, Figure 16); October 2003 is the earliest target for
458 which a full eighteen-month history exists.

459 **Table 10. Chronological division of samples by target month.**

Split	Calendar period	Usable targets	Samples
Training	Apr 2002-Mar 2020	Oct 2003-Mar 2020	168
Validation	Apr 2020-Mar 2023	Apr 2020-Mar 2023	36
Test	Apr 2023-Dec 2024	Apr 2023-Oct 2024	19



Figure 16. Chronological split and input window. Each target month t is predicted from months t - 18 to t - 1; no target appears in more than one period.

Network

Each ConvLSTM cell[18] carries a hidden map $\mathbf{H}_j$ and a memory map $\mathbf{C}_j$. For each month, the incoming fields and the preceding hidden state are concatenated into $\mathbf{Z}_j$; learned kernels $\mathbf{K}$ (bias included) are convolved with it (*), S is the logistic function and ∘ is element-wise multiplication:

$$\left(\mathbf{i}_j, \mathbf{f}_j, \mathbf{o}_j\right) = S\left(\mathbf{K}_{ifo} * \mathbf{Z}_j\right), \qquad \mathbf{G}_j = \tanh\left(\mathbf{K}_g * \mathbf{Z}_j\right), \quad (3)$$

$$\mathbf{C}_j = \mathbf{f}_j \circ \mathbf{C}_{j-1} + \mathbf{i}_j \circ \mathbf{G}_j, \qquad \mathbf{H}_j = \mathbf{o}_j \circ \tanh\left(\mathbf{C}_j\right), \quad (4)$$

in which i_j, f_j and o_j act as input, forget and output gates and G_j is the candidate update.

The network stacks two such layers with 80 and 40 hidden channels and 3 × 3 kernels with replicate padding (Figure 17). The first layer's hidden map is batch-normalised[33] at every step before entering the second. Once all eighteen months have been read, the upper hidden state is normalised, thinned by channel-wise dropout at a rate of 0.4, and decoded by a dilated[34] (rate 2) sixteen-filter 3 × 3 convolution with a rectified linear unit (ReLU) activation[35] followed by a single-filter 1 × 1 convolution. Model B has 429,873 trainable parameters and Model A 432,753.

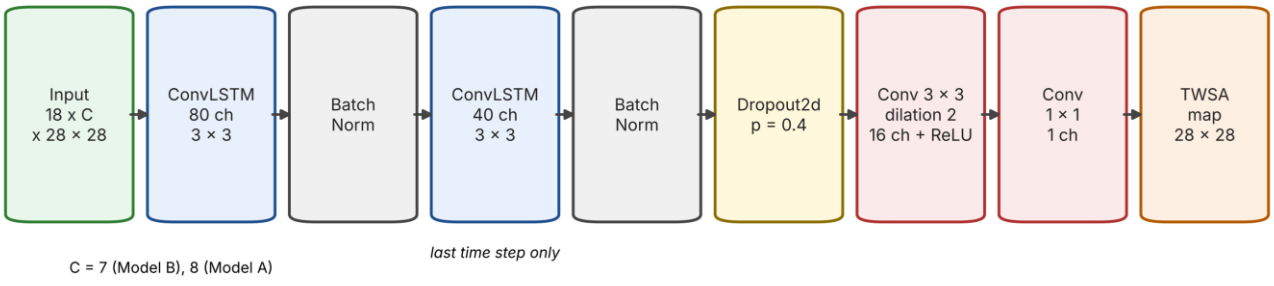


Figure 17. ConvLSTM architecture shared by both variants. Only the hidden map after the last time step reaches the output head.

Training

The loss was the Huber function[36] averaged over mask cells:

$$\mathcal{L} = \frac{\sum_k m_k H_h(p_k - y_k)}{\sum_k m_k}, \quad (5)$$

where m_k is 1 for mask cells and 0 elsewhere, and the Huber function $H_h(e)$ equals $e^2/2$ for $|e| \le h$ and $h(|e| - h/2)$ otherwise, with $h = 1$ in standardised units. Parameters were updated by Adam[37] (learning rate 5×10^{-4}, weight decay 10^{-4}); the learning rate was halved whenever five consecutive epochs brought no gain on the validation set, gradient norms were capped at 1.0, mini-batches held four samples, and training ran for 50 epochs, after which the weights that had scored best on validation were restored. Three reproducible runs (seeds 0, 1, 2) were made per variant, and their average forms the forecast we evaluate. Table 11 lists the configuration. Experiments used PyTorch 2.10[38] and scikit-learn 1.6.1[39] on a Kaggle NVIDIA T4 GPU; the complete run, including baselines and figures, took 19.4 min.

Table 11. Network and training configuration (identical for both variants apart from the input channels).

Setting	Value
Input window	18 months
Input channels	7 (Model B); 8 (Model A)
Hidden channels	80 and 40, kernel 3 × 3
Regularisation	Batch normalisation; Dropout2d p = 0.4
Output head	Conv 3 × 3 (dilation 2, 16 ch), ReLU, Conv 1 × 1
Parameters	429,873 (B); 432,753 (A)
Loss	Masked Huber, h = 1
Optimiser	Adam; learning rate 5×10^{-4}; weight decay 10^{-4}
Learning-rate rule	Halved after 5 epochs without gain
Gradient-norm cap / batch size / epochs	1.0 / 4 / 50
Weights kept	Best on validation
Seeds	0, 1, 2 (ensemble mean reported)

Reference forecasts

All references were scored on the same months and cells as the networks. Persistence predicts the most recent real observation before the target month. This assumes that the previous month's field is already available when the forecast is issued, which is not the case operationally. Monthly climatology predicts, at each cell, the mean of the observed training values for that calendar month. The Random Forest[40] is trained separately for every cell (200 trees, depth limit 15) using the five climate inputs lagged by 1, 2, 3, 6, 12 and 18 months together with a harmonic encoding of the target month; it receives neither storage values nor spatial context, so the comparison shows what the ConvLSTM gains over a non-spatial model that sees only climate; because the two models differ in several respects, it does not isolate the contribution of spatial structure alone. The forest used fixed hyperparameters that were not tuned.

Metrics and statistical tests

With y_{kt} and p_{kt} the observed and predicted storage at cell k and month t, we report RMSE, MAE, mean bias, the Pearson correlation pooled over cell-months, the median per-cell temporal correlation, and the efficiency

$$\mathrm{NSE}_k = 1 - \frac{\sum_t (p_{kt} - y_{kt})^2}{\sum_t (y_{kt} - a_k)^2}, \quad (6)$$

with a_k denoting the average observed value of cell k over the test months. We summarise it in four ways: averaged across cells, as the cell median, computed on all cell-months together, and for the state-mean series. To measure skill beyond the seasonal cycle, the training climatology $c_{k,m}$ of the target calendar month is removed from both series before correlating them:

$$\mathrm{ACC} = \mathrm{corr}\big(p_{kt} - c_{k,m(t)},\ y_{kt} - c_{k,m(t)}\big). \quad (7)$$

Improvement over a reference forecast is expressed by the skill score

$$\mathrm{MSESS} = 1 - \frac{\mathrm{MSE}_{\mathrm{model}}}{\mathrm{MSE}_{\mathrm{reference}}}, \quad (8)$$

which is positive when the model beats the reference. For paired comparisons the area-mean absolute error of each test month (in standardised units, not centimetres) was computed for both methods, giving 19 pairs. We applied the Wilcoxon signed-rank test[41] and a moving-block bootstrap[23] of the mean difference with blocks of three months and 10,000 resamples; a difference was regarded as robust only when the 95% interval excluded zero. Because cells inside one mascon repeat a single series, NSE and correlation were also computed for each distinct series; series represented by under five mask cells were labelled boundary fragments, and the summary value weights the three main mascons by their cell counts. The mean squared error was also split into the squared mean bias and the variance of the errors. For the four grid sectors, the pooled efficiency 1 − SSE/SST was computed over all cell-months of the sector, where SSE is the sum of squared errors and SST the sum of squared deviations of the observations from their mean in that sector.

Drought classes

Following the GRACE drought severity index (GRACE-DSI)[25], each value was standardised by the training-period mean c and standard deviation s of its calendar month:

$$\mathrm{DSI}_{kt} = \frac{y_{kt} - c_{k,m(t)}}{s_{k,m(t)}}, \quad (9)$$

and the same transformation was applied to predictions. Thresholds at −1.3, −0.5, 0.5 and 1.3 define five classes from severe drought to very wet. Agreement over the 4009 test cell-months (211 × 19) was measured by accuracy, macro and weighted F1 and Cohen's kappa.

Leakage safeguards

Before any training, fourteen assertions are executed against the actual files and halt execution if one fails (Table 12). They cover temporal causality, the split, the origin of every statistic, the inputs of each baseline and the expected tensor shapes. In addition, 36 unit tests cover dating, regridding, sequence construction, baselines and metrics, and a two-epoch smoke run on the real files is inspected automatically before the full experiment. A final guard compares the hash of the TerraClimate file and the run configuration with the expected values so that rehearsal outputs cannot be mistaken for final results.

Table 12. Safeguards executed automatically on the real inputs before training.

No.	Safeguard
1	Every sample uses an 18-month window
2	Inputs never extend to or beyond the forecast month
3	Altering data from the forecast month onward cannot change any input
4	Training, validation and test months are contiguous and disjoint
5	Targets are restricted to measured GRACE months
6	Z-score moments come from the training period alone
7	The storage channel of Model A lags the target and holds measured values only
8	Removing all post-2020 GRACE leaves the study mask identical
9	Model B receives seven inputs, none derived from storage
10	Model A receives exactly one extra input
11	The persistence forecast copies the most recent measured field
12	Random Forest features contain neither storage nor future climate
13	Drought-index means and spreads come from training years
14	Array dimensions match the expected layout

Data availability

The GRACE and GRACE-FO mascon data (JPL RL06.3Mv04)[27,29] and the TerraClimate data[31] are publicly available through the Google Earth Engine catalogue, and the state outline is available from DataMeet under a CC BY 4.0 licence.[28] The scripts, configuration files, exported input rasters and final-run result files that support this study are archived at Zenodo: https://doi.org/10.5281/zenodo.23491599.

References

Tapley, B. D., Bettadpur, S., Ries, J. C., Thompson, P. F. & Watkins, M. M. GRACE measurements of mass variability in the Earth system. *Science* **305**, 503–505 (2004).

Landerer, F. W. et al. Extending the global mass change data record: GRACE Follow-On instrument and science data performance. *Geophys. Res. Lett.* **47**, e2020GL088306 (2020).

Tapley, B. D. et al. Contributions of GRACE to understanding climate change. *Nat. Clim. Change* **9**, 358–369 (2019).

Rodell, M., Velicogna, I. & Famiglietti, J. S. Satellite-based estimates of groundwater depletion in India. *Nature* **460**, 999–1002 (2009).

Asoka, A., Gleeson, T., Wada, Y. & Mishra, V. Relative contribution of monsoon precipitation and pumping to changes in groundwater storage in India. *Nat. Geosci.* **10**, 109–117 (2017).

Famiglietti, J. S. The global groundwater crisis. *Nat. Clim. Change* **4**, 945–948 (2014).

Rodell, M. et al. Emerging trends in global freshwater availability. *Nature* **557**, 651–659 (2018).

564 Scanlon, B. R. et al. Global models underestimate large decadal declining and rising water
565 storage trends relative to GRACE satellite data. *Proc. Natl. Acad. Sci. USA* **115**, E1080–E1089
566 (2018).

567 Watkins, M. M., Wiese, D. N., Yuan, D.-N., Boening, C. & Landerer, F. W. Improved methods for
568 observing Earth's time variable mass distribution with GRACE using spherical cap mascons. *J.*
569 *Geophys. Res. Solid Earth* **120**, 2648–2671 (2015).

570 Wiese, D. N., Landerer, F. W. & Watkins, M. M. Quantifying and reducing leakage errors in the
571 JPL RL05M GRACE mascon solution. *Water Resour. Res.* **52**, 7490–7502 (2016).

572 Humphrey, V., Gudmundsson, L. & Seneviratne, S. I. A global reconstruction of climate-driven
573 subdecadal water storage variability. *Geophys. Res. Lett.* **44**, 2300–2309 (2017).

574 Humphrey, V. & Gudmundsson, L. GRACE-REC: A reconstruction of climate-driven water stor-
575 age changes over the last century. *Earth Syst. Sci. Data* **11**, 1153–1170 (2019).

576 Sun, A. Y. et al. Combining physically based modeling and deep learning for fusing GRACE sat-
577 ellite data: Can we learn from mismatch? *Water Resour. Res.* **55**, 1179–1195 (2019).

578 Mo, S. et al. Bayesian convolutional neural networks for predicting the terrestrial water storage
579 anomalies during GRACE and GRACE-FO gap. *J. Hydrol.* **604**, 127244 (2022).

580 Foroumandi, E., Nourani, V., Huang, J. J. & Moradkhani, H. Drought monitoring by downscaling
581 GRACE-derived terrestrial water storage anomalies: A deep learning approach. *J. Hydrol.* **616**,
582 128838 (2023).

583 Kratzert, F., Klotz, D., Brenner, C., Schulz, K. & Herrnegger, M. Rainfall–runoff modelling using
584 Long Short-Term Memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* **22**, 6005–6022 (2018).

585 Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780
586 (1997).

587 Shi, X. et al. Convolutional LSTM network: A machine learning approach for precipitation now-
588 casting. In *Advances in Neural Information Processing Systems* Vol. 28, 802–810 (2015).

589 Nash, J. E. & Sutcliffe, J. V. River flow forecasting through conceptual models part I—A discus-
590 sion of principles. *J. Hydrol.* **10**, 282–290 (1970).

591 Schaefli, B. & Gupta, H. V. Do Nash values have value? *Hydrol. Process.* **21**, 2075–2080 (2007).

592 Murphy, A. H. Skill scores based on the mean square error and their relationships to the corre-
593 lation coefficient. *Mon. Weather Rev.* **116**, 2417–2424 (1988).

594 Bergmeir, C. & Benítez, J. M. On the use of cross-validation for time series predictor evaluation.
595 *Inf. Sci.* **191**, 192–213 (2012).

596 Künsch, H. R. The jackknife and the bootstrap for general stationary observations. *Ann. Stat.*
597 **17**, 1217–1241 (1989).

598 Thomas, A. C., Reager, J. T., Famiglietti, J. S. & Rodell, M. A GRACE-based water storage deficit
599 approach for hydrological drought characterization. *Geophys. Res. Lett.* **41**, 1537–1545 (2014).

Zhao, M., A, G., Velicogna, I. & Kimball, J. S. A global gridded dataset of GRACE drought severity index for 2002–14: Comparison with PDSI and SPEI and a case study of the Australia Millennium Drought. *J. Hydrometeorol.* **18**, 2117–2129 (2017).

Kaufman, S., Rosset, S., Perlich, C. & Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data* **6**, 15 (2012).

Google Earth Engine Data Catalog. GRACE Monthly Mass Grids Release 6.3 Version 04—Global Mascons. https://developers.google.com/earth-engine/datasets/catalog/NASA_GRACE_MASS_GRIDS_V04_MASCON (accessed 1 October 2026).

DataMeet. Maps: India administrative boundaries. https://github.com/datameet/maps (accessed 1 October 2026).

Wiese, D. N., Yuan, D.-N., Boening, C., Landerer, F. W. & Watkins, M. M. JPL GRACE and GRACE-FO Mascon Ocean, Ice, and Hydrology Equivalent Water Height JPL RL06.3Mv04. PO.DAAC, doi:10.5067/TEMSC-3MJ634 (2023).

Gorelick, N. et al. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017).

Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A. & Hegewisch, K. C. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Sci. Data* **5**, 170191 (2018).

Sutterley, T. C. gravity-toolkit: Python tools for obtaining and working with GRACE and GRACE-FO gravity data. https://github.com/tsutterley/gravity-toolkit (accessed 1 October 2026).

Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. 32nd Int. Conf. Mach. Learn.* 448–456 (2015).

Yu, F. & Koltun, V. Multi-scale context aggregation by dilated convolutions. In *4th Int. Conf. Learn. Represent.* (2016).

Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *Proc. 27th Int. Conf. Mach. Learn.* 807–814 (2010).

Huber, P. J. Robust estimation of a location parameter. *Ann. Math. Stat.* **35**, 73–101 (1964).

Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *3rd Int. Conf. Learn. Represent.* (2015).

Paszke, A. et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* Vol. 32, 8024–8035 (2019).

Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1**, 80–83 (1945).

Acknowledgements

GRACE land are available at https://grace.jpl.nasa.gov, supported by the NASA MEaSUREs Program. We thank the TerraClimate developers and the Google Earth Engine team for releasing their data openly. This research received no specific grant from any funding agency. Generative AI tools were used to polish wording and grammar; the authors designed the study, performed the computations, interpreted the findings and take responsibility for the text.

Author contributions

R.P. and R.S.G. conceived the study. R.P. developed the methodology and software, curated the data, performed the formal analysis and investigation, prepared the figures and wrote the original draft. R.P., R.S.G. and A.T.M. validated the results. R.S.G. and A.T.M. reviewed and edited the manuscript, and A.T.M. and R.S.G. supervised the work. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.T.M. (tm_amirthalakshmi@ch.amrita.edu).