Supplementary Information for

Geophysics-informed groundwater mapping reveals slow-recovery zones after extreme drought

Zhengyang Fang[1,2†], Zhuo Liu[1,3†], Deshan Feng[1], Yongfei Wang[1,4], Rongwen Guo[1*], Hang Chen[2*]

[1]School of Geosciences and Info-physics, Central South University, Changsha, 410083, China.
[2]School of Earth, Environment, and Sustainability, University of Iowa, Iowa City, IA, 52245, USA.
[3]Doerr School of Sustainability, Department of Earth and Planetary Sciences, Stanford University, Stanford, California, 94305, USA.
[4]Department of Mathematical Sciences, Tsinghua University, Beijing, 100084, China.

*Corresponding author(s). E-mail(s): rongwenguo@csu.edu.cn; hchen117@uiowa.edu;
†These authors contributed equally to this work.

The PDF file includes:

- Materials and Methods
- Figs. S1 to S27
- Tables S1 to S3
- References

Materials and Methods

S1.1 Study region and data sources

The study region covers the Mississippi River Valley alluvial aquifer (MRVA) within the Mississippi Alluvial Plain (MAP). It is a heavily irrigated alluvial aquifer system in the lower Mississippi Valley. We used the USGS MAP extent to define the study boundary. All spatial data were organized on a common 1-km Albers equal-area grid and clipped to the MRVA mask [1]. The active grid contained 87,871 cells.

Groundwater-level observations were obtained from the USGS Water Data for the Nation/National Water Information System (NWIS) [2]. We downloaded discrete field measurements, daily values and instantaneous values through the USGS Water Data API [3]. The dataset included groundwater-level and hydraulic-head parameter codes. We focused on shallow alluvial wells to represent the MRVA water-table system. We excluded wells assigned to deeper aquifers or with reported total depths greater than 300 ft. We then selected one shallow or alluvial well with sufficient monthly coverage in each active grid cell. After screening, 2,514 representative wells remained in 2,514 grid cells. These wells covered 2.86% of the active MRVA grid. They provided 31,654 monthly WTD records for 2011–2023, aggregated from 187,255 mapped site-day observations. USGS site metadata were used to convert water-level altitudes to depth below land surface.

Subsurface structure was represented with USGS AEM derivative products from the James and Minsley data release [4], which combines electrical-resistivity results from four regional AEM surveys [5–8]. The 2020 release included the first two surveys [7,8]. The 2022 update added surveys conducted from March 2018 to August 2021 and from September 2021 to January 2022 [5,6]. The product follows the 1-km resolution National Hydrogeologic Grid and contains 5-m depth and elevation slices [4,9]. We used the log10 resistivity variable from the 2022 depth-grid product and sampled

70 depth levels from 0 to 350 m below land surface. The release also includes hydrogeologic products derived from the AEM models, including near-surface confinement and connectivity. Elevation data were obtained from the USGS 3D Elevation Program (3DEP), which forms part of The National Map [10]. Slope was calculated from the elevation data.

Net infiltration and pumping were used as the primary forcings. Net infiltration was obtained from the USGS Soil-Water-Balance model for the Mississippi Embayment Regional Aquifer System, 2000–2024 [11,12]. Daily values were converted from inches to metric units and summed by month on the 1-km grid. Pumping was represented using AIWUM aquaculture and irrigation water-use estimates. AIWUM2 provided the record for 2014–2023 [13]. We used the early-period AIWUM1 product for 2011–2013 [14]. The two products were not joined directly because their estimates were not on the same scale. For each month from April to October, we used the 2014–2017 overlap period to calculate the ratio between the mean AIWUM2 and AIWUM1 fields. We smoothed these spatial ratios with an 8-grid-cell Gaussian filter and limited them to 0.2–5.0. We then applied the ratios to the 2011–2013 AIWUM1 fields. Each corrected growing-season field was then rescaled so that its regional total equalled the raw AIWUM1 total multiplied by the corresponding month-specific AIWUM2-to-AIWUM1 overlap-period ratio. For January–March and November, we used the spatial patterns from the 2014–2023 AIWUM2 climatology and scaled them to each year's corrected growing-season total. December was assigned zero pumping, consistent with the AIWUM2 record. The corrected 2011–2013 fields were then combined with the 2014–2023 AIWUM2 record. The resulting series was sampled at active MRVA grid cells.

Monthly precipitation was obtained from Daymet Version 4 Revision 1 monthly climate summaries at 1-km resolution [15]. Long-term WTD was mapped with a random-forest model following the high-resolution conterminous United States WTD

framework [16,17]. We adapted this framework to the MRVA by using long-term groundwater-level observations as anchors and adding AEM resistivity variables as predictors. The resulting spatially continuous long-term WTD reference was produced at 1-km resolution and used as the baseline surface for the monthly reconstruction. For independent validation of basin-scale groundwater-storage variability, we used CSR GRACE/GRACE-FO RL06.3 mascon terrestrial-water-storage anomalies [18,19]. River-network data were obtained from HydroRIVERS/HydroSHEDS [20].

S1.2 Graph neural-network model and training

We trained separate graph neural-network (GNN) [21] models to estimate monthly changes of WTD from sparse well observations, net infiltration, pumping and AEM resistivity. Each model was trained to predict WTD change, Δh, over a fixed prediction interval H. Because WTD is expressed as depth to water below land surface, the target was defined as

$$\Delta h(t, H) = -\left[WTD(t + H) - WTD(t)\right]$$

where H is the prediction interval in months. Positive Δh denotes groundwater-level rise, and negative Δh denotes groundwater-level decline. We trained separate models for 1-, 3- and 6-month intervals. Each sample represented one grid cell and one start month with WTD observations at both the start and target months. Start-month WTD represented the current groundwater state. The model then predicted Δh between the two months. We used a fixed spatiotemporal split. We assigned 80% of well-observed grid cells to the training-grid set and reserved the remaining 20% for spatial testing. Within the training-grid set, samples from 2011–2019 were used for training, samples from 2020–2021 for temporal validation, and samples from 2022–2023 for temporal testing. The spatial test used 2011–2019 samples from the held-out grid cells. For the 1-month model, Δh was calculated from WTD estimates at the

start and target months. A monthly WTD estimate based on only one or a few daily observations was less reliable. It could also make the estimated 1-month change less reliable. We therefore retained validation and test samples only when both months had observations on at least 10 days. During training, each sample was weighted by the lower observation-day count across its two endpoints. The 1-month model used 2,933 training samples, 619 validation samples, 811 temporal-test samples and 410 spatial-test samples. The corresponding sample numbers for the 3-month model were 3,059, 865, 1,085 and 687. For the 6-month model, they were 8,485, 2,504, 3,732 and 2,274, respectively.

The input parameters were grouped into Hydrologic Forcing, Terrain Property, Forecast Time and AEM Resistivity (Fig. 1c). Net infiltration was derived from monthly Soil-Water-Balance model output. AIWUM pumping was converted to an equivalent water depth in millimeters. We applied the natural-logarithmic transformation $\ln(x+1)$ before model input to reduce the influence of highly skewed pumping values. Subsurface structure was described by 30 AEM resistivity layers in the upper 150 m. Terrain attributes included mean elevation and mean slope. Prediction timing was represented by the interval length and sine and cosine terms for the start month. All models used cumulative net infiltration and pumping over the prediction interval. For the 3- and 6-month models, future forcing sequences from $t+1$ to $t+H$ were additionally encoded so that the model could represent the recharge and pumping within the prediction interval.

The model used separate encoders for AEM resistivity, terrain attributes, prediction time and forcing summaries (Fig. 1c). The AEM profile was encoded with a multilayer perceptron. Elevation and slope were encoded as terrain features. Prediction interval and seasonal timing were encoded separately. Net infiltration and pumping summaries were encoded as forcing features. For the 3- and 6-month models, a gated recurrent unit encoded the forcing sequence before it was combined with the other

node features. The combined features were passed through weighted graph-convolution layers. The prediction head combined the graph-derived feature at the focal grid cell with standardized start-month WTD to predict Δh. Graph edges linked nearby grid cells and gave greater weight to pairs with similar AEM resistivity profiles. All graphs were made undirected. Cells separated by large resistivity contrasts had lower edge weights.

Figs. S4, S5 and Table S1 show model performance in independent temporal and spatial test sets for each prediction interval [22]. The models retained skill in both tests, indicating that they captured WTD changes in months and locations withheld from training. The 1-month model achieved RMSE values of 0.314 m in the temporal test and 0.411 m in the spatial test. For the 3-month model, the Pearson correlation coefficient was $r = 0.773$ and NSE = 0.575 [23] in the temporal test, and $r = 0.663$ and NSE = 0.430 in the spatial test. The 6-month model showed the strongest agreement with observed changes, with $r = 0.819$ and NSE = 0.663 in the temporal test, and $r = 0.812$ and NSE = 0.638 in the spatial test. Correlation and NSE increased with prediction interval in both temporal and spatial tests. 1-month WTD change is derived from WTD estimates at two monthly endpoints. Uncertainty in either monthly estimate can therefore strongly affect the estimated change. Over 3- and 6-month intervals, the cumulative WTD response is larger relative to this endpoint uncertainty. Seasonal patterns in net infiltration and pumping are also more clearly represented.

S1.3 Monthly WTD reconstruction

We reconstructed monthly WTD from January 2011 to December 2023 at each active 1-km grid cell. The workflow is shown in Fig. S6. It had three stages. First, we combined well observations with the long-term WTD map to create annual anchor maps. Second, we propagated WTD between anchors with the 1-, 3- and 6-month GNN models. Third, we corrected monthly fields with available well observations.

The first step is the reconstruction of anchor maps. The long-term WTD map provided the spatial baseline for the reconstruction [16,17]. Let B_i denote the baseline WTD at grid cell i. We converted each monthly well observation to an anomaly:

$$a_{i,t}^{obs} = WTD_{i,t}^{obs} - B_i.$$

Positive anomalies indicate deeper WTD than the long-term baseline. Negative anomalies indicate shallower WTD. We used 14 anchor months. These included January of every year from 2011 to 2023 and December 2023. The January anchors divided the record into regular annual intervals. The December 2023 anchor closed the final interval and avoided extrapolation beyond the study period. The upper part of Fig. S6 shows how each anchor map was constructed. We first calculated a local anomaly and a support score for each observed grid cell. The local anomaly was the median WTD anomaly across the observation record. The support score increased with the number of observed months and decreased when the observed anomalies had a large temporal spread. The local anomaly was reduced when support was low. We then estimated a regional monthly anomaly. For each observed cell, we calculated the departure of the monthly anomaly from its cell-specific median. We used the median departure across all observed cells as the regional signal. Missing months were linearly interpolated. The series was smoothed with a centered 3-month moving mean and centered across the full record. Some cells had no effective local support. This included cells without observations and cells with only one observed month. For these cells, we estimated anchor-month anomalies from a weighted regression between baseline WTD and observed anomalies. The regression was fitted when at least 10 supported cells were observed in that anchor month. Otherwise, we used a regression fitted across the full record. We combined the local anomaly, regional anomaly and regression estimates to create a provisional anchor field. Direct observations had the highest weight. We

used graph-based spatial interpolation to extend the anchor field across the MRVA. This allowed WTD anomalies to vary while preserving direct observations exactly.

The second step is the monthly propagation as shown in the middle and lower parts of Fig. S6, which is the reconstruction between two anchor maps. The GNN models predicted groundwater-level change, $\Delta h(t, H)$, over prediction intervals of H = 1, 3 and 6 months. We therefore updated the WTD anomaly as:

$$\widehat{a}_{i,t+H} = \widehat{a}_{i,t} - \widehat{\Delta h}_{i,t,H}.$$

We averaged predictions from the five trained seeds for each prediction interval. The GNN models used absolute WTD as input. We therefore added the baseline WTD to the current anomaly before each prediction. The reconstruction did not use a fixed interval length. Instead, it selected a sequence of 1-, 3- and 6-month predictions for each gap between anchors. The selected path minimized the sum of interval-specific validation costs. For each prediction interval, the validation cost was defined as the validation RMSE divided by the validation NSE, with NSE bounded below by 0.05. For example, the initial least-cost path for a 12-month anchor gap was 6+6 months. Intervals that exceeded the endpoint-mismatch thresholds were subdivided and reconstructed recursively. The propagated state at the end of an interval was compared with the next anchor map. The anchor was used as a soft constraint. Its weight increased when more observations were available in the anchor month. Its weight decreased when the propagated state and anchor map differed strongly. The endpoint difference was distributed through the interval with a cubic taper. The correction was zero at the first anchor and reached its full value at the next anchor. We tested each interval before applying the endpoint correction. An interval was divided into shorter segments if the uncorrected mismatch at the next anchor exceeded 2.0 m RMSE or 3.0 m at the 95th percentile of absolute error. The shorter segments were then reconstructed separately.

The third step used observed wells to constrain local WTD patterns and the regional groundwater-storage anomaly. At cells with observations in at least two months, observed WTD values were retained exactly in those months. Cells with one observed month received a small quality-dependent adjustment rather than a hard constraint. We also adjusted the reconstructed series between successive well measurements. For each observed cell, we calculated a linear bridge between two observed WTD values. The bridge had less influence when the observation gap was long or observation quality was low. We then calculated monthly residuals between reconstructed and observed WTD. This step was applied only when observations were available in at least 20 grid cells. The interpolated residual field was clipped to ± 2.5 m before applying a gain of 0.70, which bounded the applied increment at ± 1.75 m.

Direct observations were retained exactly after residual diffusion. The same well network was used to constrain the regional groundwater-storage anomaly. We converted reconstructed WTD anomalies to a regional groundwater-storage anomaly using a connectivity-informed specific-yield field [4,24], $\mathrm{S}_{y,i}$. Specific-yield values ranged from 0.08 to 0.26 across the mapped connectivity classes (Fig. S3c). Cells without a mapped class used 0.18.

$$S_t = -100 \frac{1}{N} \sum_{i=1}^{N} S_{y,i} \left(WTD_{i,t} - \overline{WTD_i} \right),$$

where S_t is expressed in centimeters, N is the number of active grid cells, and $\overline{WTD_i}$ is the mean reconstructed WTD for cell i over the reference period.

We applied one support-weighted shift each month to unobserved cells so that the reconstructed regional groundwater-storage anomaly matched the well-derived storage target. The shift decreased as local well support increased. Direct observations were retained.

S1.4 Reconstruction validation and uncertainty

We assessed the final reconstruction in two ways. First, we quantified uncertainty in the monthly WTD field. Second, we compared regional groundwater-storage anomalies derived from the reconstruction with GRACE data. The temporal and spatial holdout tests of the GNN models are reported in Section S1.2. Those tests evaluate predictive skill before reconstruction. The analyses below evaluate the reconstructed monthly field after it has been constrained by well observations and anchor maps.

We estimated nominal 75% prediction intervals (PI75) with normalized split-conformal calibration [25]. For each prediction interval H, the five GNN seeds provided an ensemble mean prediction, $\widehat{\Delta h}_{H,j}$, and an ensemble standard deviation, $\sigma_{H,j}$ [26]. We calculated normalized residuals in the temporal-validation set:

$$s_{H,j} = \frac{\left|\Delta h_{H,j} - \widehat{\Delta h}_{H,j}\right|}{\max\left(\sigma_{H,j}, 10^{-6}\right)}$$

where $\Delta h_{H,j}$ is the observed WTD change. We defined q_H as the 75th conformal percentile of these residuals. The PI75 radius for each prediction was

$$R_{H,i} = q_H \sigma_{H,i}.$$

The PI75 was therefore $\widehat{\Delta h}_{H,i} \pm R_{H,i}$. The fitted values of q_H were 4.433, 4.770 and 4.494 for the 1-, 3- and 6-month models, respectively. The interval width varied with disagreement among the five GNN predictions.

Some reconstruction segments used more than one prediction step. We combined the PI75 radii from these steps with a discounted root-sum-of-squares rule, using γ=0.55. At months with well observations, we used the 1-month model uncertainty. Fig. S7 illustrates three grid cells that contained 154, 21 and 1 monthly observations during 2011–2023. Their mean PI75 radii were 0.43, 0.76 and 1.18 m, respectively. At the densely observed cell, the reconstructed series is tightly constrained by the monthly

observations and has a narrow interval. At the cell with 21 observations, the interval is wider because the reconstruction relies more often on model propagation between observations. At the cell with one observation, the series is determined mainly by the long-term baseline, interval predictions, anchor maps and regional well constraints.

Fig. S8a maps the mean PI75 radius for each grid cell across the 156 reconstructed months. The median PI75 radius across all grid-cell months was 0.73 m. The interquartile range was 0.47–1.22 m, and the 95th percentile was 2.64 m. The median of the mean radius for each grid cell was 0.85 m. Its 5th and 95th percentiles were 0.44 and 2.07 m, respectively. Fig. S8b shows WTD reconstruction confidence derived from the mean PI75 radius at each 1-km grid cell across the 156 reconstructed months. We assigned high confidence to cells with a mean radius of 0.70 m or less, and medium confidence to cells with radii from 0.70 to 1.50 m. Cells with radii greater than 1.50 m were assigned low confidence. These classes summarize uncertainty in reconstructed WTD. Based on the mean PI75 radius, 86.2% of the MRVA had high or medium WTD reconstruction confidence. The representative time series showed wider PI75 intervals where direct well observations were sparse.

We used GRACE as an external regional-scale comparison (Fig. S9) [18,19] and converted reconstructed WTD anomalies to groundwater-storage anomalies with the connectivity-informed specific-yield field described in Section S1.3. Each reconstructed cell was referenced to its mean WTD during valid GRACE months. GRACE measures total water storage, not groundwater storage alone. It also has a much coarser spatial resolution than the 1-km reconstruction. The comparison therefore tests whether the reconstructed field captures the characteristics of regional storage change. For each subregion, we calculated a GRACE mean weighted by the number of assigned 1-km MRVA cells in each GRACE grid cell. We calculated Pearson's correlation coefficient r for valid GRACE months. The reconstructed anomaly tracked the GRACE record at the MRVA scale ($r = 0.91$). Correlations were positive in Boeuf ($r = 0.53$),

Cache–Grand Prairie ($r = 0.56$), Delta ($r = 0.77$) and St. Francis ($r = 0.75$). Agreement was strong at the scale resolved by GRACE. The MRVA-wide correlation shows that the reconstruction captured the dominant seasonal and multi-year variations in regional storage. The positive correlations across the four subregions (0.53–0.77) show that temporal agreement persisted below the basin scale, although this comparison is limited by the coarse GRACE footprint and its inclusion of non-groundwater storage.

S1.5 Drought-response metrics

Groundwater droughts can differ in the magnitude and rate of water-level decline, as well as in the timing and completeness of recovery [27–29]. We quantified drought response from reconstructed monthly WTD using the 2012 drought as a common event [30]. Figs. S10–S12 show precipitation, net infiltration and pumping during the three drought periods and across 2011–2023. These variables provide the hydroclimatic and water-use context, but they were not used to calculate the response metrics or assign response classes. The metrics describe WTD before, during and after the 2012 drought. They comprise two decline metrics, maximum decline and decline rate, and three recovery metrics, early recovery ratio, half-recovery time and long-term recovery ratio (Fig. S13). Together, they describe the magnitude and rate of WTD decline, short-term recovery and recovery by 2019. We first defined the reference states used to calculate these metrics.

We defined the pre-drought state as the shallowest WTD from January to April 2012 and the drought peak as the deepest WTD from May to October 2012:

$$W_{pre} = \min\left(W_{Jan-Apr\,2012}\right)$$

$$W_{peak} = \max\left(W_{May-Oct\,2012}\right)$$

The shallowest WTD from November 2012 to April 2013 is defined as early recovery state before the next growing-season:

$$W_{early} = \min\left(W_{Nov\,2012-Apr\,2013}\right)$$

Therefore, maximum decline was the difference between these two values:

$$\Delta W_{\max} = W_{peak} - W_{pre}.$$

Decline rate measures how quickly the water table deepened. We divided maximum decline by the number of months between the pre-drought minimum and the drought peak:

$$R_{down} = \frac{\Delta W_{\max}}{\max\left(t_{peak} - t_{pre}, 1\right)}$$

And the early recovery ratio measures the fraction of the 2012 decline recovered:

$$RR_{early} = \frac{W_{peak} - W_{early}}{\Delta W_{\max}}$$

A value of 1 indicates recovery to the pre-drought reference. A value below 1 indicates partial recovery. Values above 1 indicate that WTD became shallower than the pre-drought reference during the early-recovery window. Half-recovery time, $T50$, measures the time required to recover half of the maximum decline. The recovery threshold was

$$W_{50} = W_{peak} - 0.5\Delta W_{\max}$$

We identified the first month after the drought peak when $WTD \leq W_{50}$. T_{50} was the number of months between the drought peak and this first recovery month. It was recorded as missing when the threshold was not reached by December 2023. Long-term recovery ratio describes the later recovery state. We used the mean WTD in 2011 as

the baseline and the deepest WTD during the full 2012 calendar year as the maximum annual deficit:

$$D_{2012} = W_{peak,2012} - \overline{W}_{2011}$$

The residual deficit in 2019 was

$$D^{res}_{2019} = \overline{W}_{2019} - \overline{W}_{2011}$$

The long-term recovery ratio was calculated as

$$RR_{2019} = \frac{D_{2012} - D^{res}_{2019}}{D_{2012}}$$

A value of 1 indicates that the mean WTD in 2019 matched the 2011 baseline. Values below 1 indicate a remaining deficit. Values above 1 indicate that the 2019 mean WTD was shallower than the 2011 baseline. We retained the raw recovery ratios in the metrics table. We treated short-term metrics as unreliable when maximum decline was below 0.30 m. In these locations, maximum decline, decline rate, early recovery ratio and half-recovery time were set to missing. We treated long-term recovery ratio as unreliable when the 2012 deficit was below 0.10 m.

Fig. S14 shows that the five metrics describe a broad and continuous range of drought responses across the MRVA. Complete values were available for 83,913 1-km grid cells, representing 95.5% of the active grid. The median maximum decline was 1.49 m, the median decline rate was 0.27 m/month, the median early recovery ratio was 0.73, the median half-recovery time was 4 months, and the median long-term recovery ratio was 1.36. Maximum decline and decline rate were moderately related (Spearman's ρ=0.57) which shows that a large decline did not always develop rapidly. Early recovery ratio and half-recovery time were strongly inversely related ($\rho = -0.70$). Long-term recovery ratio had weaker relationships with the other metrics ($|\rho| \leq 0.38$), so it added information beyond the short-term response.

The first two principal components explained 74.1% of the variation in the five metrics [31]. The first component described a gradient from large declines with rapid early recovery and short half-recovery times to weaker or more persistent responses. The second component was associated with rapid decline, low early recovery, long half-recovery time and lower long-term recovery.

S1.6 Response classification

We used a physically constrained hard-classification approach. Each grid was compared with predefined response prototypes in a common percentile-rank space. The three groups were fast recovery, slow recovery and buffered zones. A grid was eligible for classification when maximum decline, decline rate, early recovery ratio and half-recovery time were all available. This retained 87,647 of the 87,871 active MRVA grid cells (99.75%). Long-term recovery ratio was not required because it is undefined where the 2012 deficit was too small. A missing long-term recovery ratio was assigned a neutral rank during classification. This allowed locations with a clear event-scale response to remain in the analysis. We transformed each metric to an empirical percentile rank between 0 and 1. Maximum decline and decline rate were ranked directly. Before ranking, early recovery ratio was limited to 0–2, half-recovery time to 1–72 months, and long-term recovery ratio to −1–5. These limits reduced the influence of extreme recovery values on the distance calculation. The original metrics table retained the raw values. We represented incomplete long-term recovery as a long-term deficit:

$$r_{deficit,2019} = 1 - r\left(RR_{2019}\right)$$

where r is the percentile rank. A larger value therefore indicates weaker recovery by 2019. The three response prototypes were specified from the WTD responses. Fast-recovery zones were expected to show substantial drought deepening followed by rapid

recovery. Slow-recovery zones were expected to show low early recovery, long half-recovery times and a greater residual deficit by 2019. Buffered locations were expected to show weak WTD decline and a small long-term deficit. Table S2 lists the prototype coordinates and metric weights used for the classification.

Values in Supplementary Table S2 are target percentile ranks, and values in parentheses are metric weights. The weights sum to 1 within each response group. For each grid location, we calculated the weighted Manhattan distance to each prototype:

$$d_k = \sum_{j=1}^{5} w_{k,j} \left| r_j - p_{k,j} \right|,$$

where d_k is the distance to response group k, r_j is the percentile rank of metric j, $p_{k,j}$ is the corresponding prototype value, and $w_{k,j}$ is its weight. Each location was assigned to the group with the smallest distance.

We applied an additional Buffered gate to preserve the physical definition of weak drought response. A location was directly assigned to Buffered when its maximum-decline rank was 0.22 or lower, its decline-rate rank was 0.28 or lower, and its 2019-deficit rank was 0.40 or lower. This rule prevented locations with weak decline and little residual deficit from being assigned to Slow recovery because of unstable recovery ratios. Class-specific distributions of the five drought-response metrics are summarized in Table S3.

We evaluated the stability of the three-class assignment under reconstruction uncertainty (Fig. S15). We generated 64 independent perturbations of the reconstructed WTD field. Within each interval between successive anchors, we added smooth WTD perturbations scaled by the PI75 uncertainty radius. We recalculated these response metrics and repeated the same three-class assignment for each perturbed reconstruction. Fig. S15a shows the classes derived from the unperturbed reconstruction. For each location, Fig. S15b shows the response group that occurred most often across the 64 perturbed reconstructions. Fig. S15c shows the proportion

of draws assigned to this most frequent group. This proportion is the classification confidence. The most frequent perturbed class matched the original class at 82.4% of classified locations. Agreement was 91.5% for Fast recovery, 83.5% for Buffered and 71.8% for Slow recovery. These results indicate that the main spatial pattern of the three response groups is stable under the propagated WTD uncertainty.

S1.7 Resistivity and pumping analyses by response class

We compared electrical resistivity, pumping and precipitation across the three response classes after classification. These variables were not used to calculate the drought-response metrics or assign the response classes.

For the subsurface comparison, we extracted the AEM resistivity value at each of the 70 depth levels from 0 to 350 m below land surface. We calculated the median resistivity of each response class at every depth level to obtain the class profiles shown in Fig. S16a. We also divided the profile into ten depth intervals. Within each interval, all finite resistivity values were assigned to five ranges: 0–5, 5–10, 10–20, 20–30 and $\geq$30 Ω·m. Their proportions describe the vertical resistivity composition of each class (Fig. S16b). Across all three response classes, resistivity decreased with depth. This common trend probably reflects several overlapping controls, including sediment texture, saturation, pore-water chemistry and sediment compaction. Near-surface resistivity may also be elevated in coarse or partly unsaturated sediments. Because these effects cannot be separated from the AEM data alone, we interpreted the class differences relative to this shared pattern variation with depth.

Slow-recovery zones retained the highest median resistivity through most of the upper 250 m, consistent with a thicker and more continuous resistive sequence. Median resistivity in the 0–15 m interval was 45.8 Ω·m, compared with 39.6 Ω·m in Buffered zones and 35.1 Ω·m in Fast-recovery zones. This ordering persisted at intermediate

depths. At 100–150 m, median resistivity was 15.2, 12.6 and 9.3 Ω·m for Slow recovery, Buffered and Fast recovery, respectively. Fast-recovery zones showed the sharpest transition from a resistive shallow interval to lower-resistivity material below approximately 50 m. Buffered zones also showed a marked decrease in resistivity between 50 and 100 m relative to the overlying sediments. This lower-resistivity interval is consistent with finer-grained, less permeable material that may restrict vertical hydraulic connectivity and locally separate the shallow and intermediate-to-deep aquifer units. This layered structure was most evident in the representative Buffered sections shown in Fig. S19, whereas the class-wide profile showed greater structural variability. The profiles converged below approximately 300 m, where median resistivity was generally less than 5 Ω·m. Within the MAP, higher resistivity is generally consistent with coarser, more sand-rich and hydraulically conductive material, whereas lower resistivity commonly corresponds to finer, more clay-rich sediment [32–35]. The depth-band compositions indicate that the class differences arose from changes in both the proportion and vertical persistence of these resistivity ranges.

For each response class, we calculated the pumping volume at each 1-km grid cell for May–September 2012 (Fig. S17a). We also summarized four aspects of pumping pressure: total pumping during May–October 2012, maximum monthly pumping during May–September 2012, mean May–September pumping over 2011–2023 and the proportion of grid cells with pumping during May–September 2012. The latter four quantities were divided by the corresponding Fast-recovery value for the relative comparisons in Fig. S17b. Pumping differed consistently among the response classes. Median May–October 2012 pumping was 150,896 m^3 in Slow-recovery zones, compared with 95,258 m^3 in Fast-recovery zones and 112,102 m^3 in Buffered zones. Slow-recovery pumping was therefore 1.58 times the Fast-recovery value and 1.35 times the Buffered value. The same order was present for peak monthly pumping, with medians of 51,066, 32,649 and 38,336 m^3, respectively. Slow-recovery zones also had the highest long-term

irrigation-season pumping, at 126,615 m^3, compared with 83,654 m^3 for Fast recovery and 115,009 m^3 for Buffered. Pumping occurred in 78.9% of Slow-recovery grid cells, 75.2% of Buffered grid cells and 68.3% of Fast-recovery grid cells. The elevated withdrawal in Slow-recovery zones was therefore present in both the 2012 drought and the longer pumping record.

We also compared precipitation among the classes to assess whether the response differences could be explained by climatic input alone (Fig. S18). Monthly precipitation was summed for 2011–2023, January–April 2012, May–October 2012 and November 2012–April 2013. Fast-recovery and Buffered zones received similar precipitation during the 2012 drought, with median totals of 609.5 and 597.9 mm, respectively. Their early-recovery totals were also similar, at 862.5 and 854.0 mm. Slow-recovery zones received 495.3 mm during the drought and 698.6 mm during early recovery, about 18% less than the other two classes. Precipitation therefore contributed to the climatic setting of the response, but it did not explain the class separation on its own. Fast-recovery and Buffered zones received nearly the same precipitation but showed different WTD responses. The more persistent class differences occurred in pumping intensity and resistivity structure.

Fig. 3 presents one representative well from each response class. Fig. S19 extends this comparison with representative wells from different locations within each class. Each supplementary site was at least 20 km from the MRVA boundary and contained observations in at least 80 months spanning ten years. Fast-recovery examples showed a strong resistivity contrast between the shallow sedimentary interval and the underlying sedimentary sequence. Slow-recovery examples contained a resistive interval at least 100 m thick, with resistivity above 20 Ω·m. Buffered examples contained resistive shallow and deeper hydrostratigraphic intervals separated by a lower-resistivity middle interval. This middle interval is consistent with finer-grained sediments that may act as a local confining unit and restrict vertical hydraulic connectivity. Precipitation

contributed to the climatic setting of the 2012 response but did not fully account for the separation among the three classes. More persistent differences occurred in pumping intensity and resistivity structure.

S1.8 Slow-recovery probability model

The class comparisons showed that slow-recovery zones were associated with distinct resistivity structures and stronger pumping. We tested whether these two properties were sufficient to identify Slow recovery zones without using the WTD response metrics that defined the classes. Slow recovery was treated as the positive class. Fast recovery and Buffered were combined as the comparison class. The model output was the probability of Slow recovery membership

$$P_{S,i} = \Pr\left(y_i = 1 \mid \mathbf{r}_i, \mathbf{p}_i\right),$$

where $y_i = 1$ denotes Slow recovery, $\mathbf{r}_i$ represents the AEM resistivity structure, and $\mathbf{p}_i$ represents pumping. The analysis included 87,549 grid locations, of which 35.2% belonged to the Slow-recovery class. The five drought-response metrics, reconstructed WTD, precipitation, net infiltration and spatial coordinates were not included as predictors. This separation avoided using the same WTD information to define and predict the response classes. The predictor set contained 77 variables. Subsurface structure was represented by the $\log_{10}$-transformed AEM resistivity at 70 depth levels from 0 to 350 m below land surface. We also included the mean, standard deviation and range of each complete resistivity profile. Pumping was represented by four variables: total pumping during May–October 2012, mean growing-season pumping during 2011–2023, mean annual pumping during 2011–2023 and peak monthly pumping during 2011–2023. Pumping variables were transformed with $\ln(x + 1)$ to reduce their positive skew. Missing predictor values were replaced with the median of the corresponding variable.

We used an ExtraTrees classifier to represent nonlinear relations among resistivity at different depths, pumping and response-class membership [36]. The model contained 500 trees. Each terminal node contained at least ten training locations, and each split considered 75% of the available predictors. We did not include manually defined interactions between pumping and resistivity. The model inferred these relations through the tree structure. Model performance was evaluated with fivefold spatial-block cross-validation [22]. The MRVA grid was divided into 50-km spatial blocks, and all locations within a block were assigned to the same fold. Each fold was withheld in turn, and the remaining folds were used for model training. This procedure produced an out-of-fold probability for every modelled location and prevented nearby locations from being divided between the training and test sets. The combined model distinguished Slow-recovery zones from the other response classes under spatial cross-validation (Fig. S20). The pooled out-of-fold ROC-AUC was 0.809 [37]. Average precision was 0.694 [38], compared with a Slow-recovery prevalence of 0.352. The Brier score was 0.167, below the prevalence-only value of approximately 0.228 [39]. The reliability curve remained close to the 1:1 line across most probability bins, although no post-hoc probability calibration was applied [40]. Performance varied among the five spatial folds. ROC-AUC ranged from 0.751 to 0.914, average precision from 0.633 to 0.790 and Brier score from 0.111 to 0.194 (Fig. S20f). The model therefore retained useful discrimination across spatially separated areas.

Ablation experiments separated the information provided by subsurface structure and pumping. The AEM-only model achieved a ROC-AUC of 0.787, an average precision of 0.663 and a Brier score of 0.176. Pumping alone provided limited discrimination, with corresponding values of 0.579, 0.430 and 0.226. Combining AEM and pumping improved all three measures, increasing ROC-AUC to 0.809 and average precision to 0.694 while reducing the Brier score to 0.167. Pumping provided complementary information, although most of the predictive signal was contained in the resistivity

structure. We quantified this pattern with grouped permutation importance calculated within the withheld spatial folds [41]. All predictors in a group were permuted together, and importance was measured from the resulting decrease in ROC-AUC. The AEM resistivity profile accounted for a median 81.5% of the normalized permutation importance, the three profile summaries accounted for 2.9%, and pumping accounted for 15.4% (Fig. S20a).

The depth-group analysis showed that useful resistivity information extended through the upper 300 m (Fig. S20d). Joint permutation of the 150–300 m layers produced the largest median decrease in ROC-AUC, at 0.108. The decreases were 0.105 for the 50–150 m interval and 0.090 for the upper 50 m. Permuting the 300–350 m layers produced almost no additional loss in performance. Pumping reduced ROC-AUC by 0.053 when permuted as a group.

After spatial validation, the model was fitted to all available labelled locations and used to map P_S across the MRVA (Fig. S15d). High probabilities formed broad zones in the northern and western parts of the aquifer, together with narrower corridors in the central MRVA. Lower probabilities were more common toward the south and southeast. After refitting the model to all labelled locations, the highest-probability area, constrained to the observed Slow-recovery prevalence of 35.2%, contained 93.9% of the mapped Slow-recovery locations. This value describes agreement with the labels used for model fitting. Under 50-km spatial-block cross-validation, the corresponding out-of-fold recall and precision were both 64.7%, providing a more conservative estimate of performance in spatially independent areas.

S1.9 Pumping-reduction experiments

The preceding analyses identified spatial differences in drought recovery and the subsurface and pumping conditions associated with Slow-recovery zones. We next examined how groundwater conditions changed when pumping was reduced. For each

fitted model, the reduced-pumping and historical-pumping simulations started from the same WTD field. Pumping was the only input that differed. The WTD response was calculated as

$$\delta D_{i,s,k}(t) = D_{i,s,0}(t) - D_{i,s,k}(t)$$

where $D_{i,s,0}$ is WTD under historical pumping and $D_{i,s,k}$ is WTD under pumping strategy k. Positive δD indicates a shallower water table under reduced pumping. The observation-corrected reconstruction was retained as the groundwater history, and the simulated WTD response was applied as

$$D^{managed}_{i,s,k}(t) = D^{recon}_{i}(t) - \delta D_{i,s,k}(t)$$

The simulations used the 1-, 3- and 6- networks described in Section S1.2 and the propagation procedure described in Section S1.3. Well assimilation, regional storage adjustment and residual diffusion were not applied during the pumping-reduction simulations. Pumping was reduced before the transformed pumping variables were recalculated. Precipitation, net infiltration and all other predictors were unchanged.

Two seasonal intervention experiments were conducted from May through October in 2012 and 2013 to compare pumping reductions during the drought and subsequent recovery period. Historical pumping resumed after each intervention. Separately, a diagnostic unit-response experiment reduced pumping by 20% in each management unit individually throughout 2012. This experiment was used to derive local recovery, persistence and management leverage, as described in Section S1.10.

We converted the simulated WTD improvement to water-table-derived storage benefit using the same spatially varying specific-yield field used in the groundwater reconstruction:

$$\Delta S_{(s,k)}(t) = \sum_i A_i S_{y,i} \delta D_{(i,s,k)}(t).$$

where A_i is Cell area and $S_{y,i}$ is the specific yield used in the groundwater reconstruction. The same S_y field was used for every pumping strategy. Storage benefit was evaluated at the end of each intervention and after historical pumping resumed. Storage benefit and benefit per unit pumping reduction were evaluated for the seasonal interventions.

S1.10 Spatial response to reduced pumping

Pumping was reduced by 20% in each management unit individually from January through December 2012, and historical pumping resumed in January 2013. Local recovery, persistence and management leverage were calculated at 1-km resolution from the corresponding water-table response. For cell j, model seed s and unit m, the WTD response at month H was

$$\delta D_{j,s}(H) = D^0_{j,s}(H) - D^r_{j,s}(H).$$

where D^0 and D^r denote WTD under historical and reduced pumping, respectively. Positive δD indicates a shallower water table.

Local recovery was defined as the additional water-table improvement in December 2012:

$$R_{j,s} = \delta D_{j,s}(12).$$

Persistence measured the fraction of this response remaining in December 2013:

$$P_j = \frac{\delta D_{j,s}(24)}{\delta D_{j,s}(12)}, \quad \delta D_{j,s}(12) > 0.001\ m.$$

Values below 1 indicate decay of the initial benefit, while values above 1 indicate that the benefit increased during 2013. The Persistence map showed a pronounced spatial pattern along major river corridors, where the benefit of reduced pumping tended to decay more rapidly after historical pumping resumed (Fig. S21b). We examined whether this pattern was related to river proximity and stream–aquifer connectivity. Persistence increased with distance from major rivers and was also higher where streambed vertical electrical conductance (VIC) was larger; lower VIC indicates greater potential hydraulic connection between the stream and aquifer (Fig. S22a,b). These relations remained positive when river distance and VIC were evaluated together after removing broad differences among management units (Fig. S22c). The pattern is consistent with faster hydraulic adjustment in areas more strongly connected to surface water, which would allow the water-table benefit created by temporary pumping reductions to dissipate more quickly after historical pumping resumes.

Management leverage normalized Local recovery by the equivalent pumping reduction:

$$L_{j,s} = \frac{\delta D_{j,s}(12)}{\frac{\Delta V_m}{A_m}}.$$

where ΔV_m and A_m are the conserved pumping volume and area of the associated management unit. L is dimensionless and represents WTD response per unit pumping reduction. The metrics were calculated for each of the five fitted GNN models and averaged for mapping and subsequent analyses.

S1.11 Spatial allocation of pumping reductions

The spatial differences identified in the 2012 experiments were used to construct alternative pumping allocations for the 2013 recovery period. We compared three approaches. Uniform allocation applied the same proportional reduction across all pumping locations. High-pumping allocation gave priority to locations with greater

historical pumping during the intervention period. Leverage-guided allocation ranked locations using the management leverage derived from the 2012 experiments.

Leverage-guided allocation reduced pumping according to the 2012 management-leverage ranking. For the main comparison, the total reduction was fixed at 20% of regional 2012 pumping, so that the 2012 and 2013 interventions conserved the same pumping volume. Uniform and High-pumping allocations were matched to the same total reduction, with no individual location reduced by more than 50%. Additional reduction levels from 0 to 30% were also tested, and the distribution of the 20% reduction among the three recovery zones is shown in Fig. 4c.

S1.12 Evaluation of regional storage and local recovery

The pumping allocations described were evaluated against two related management outcomes: regional storage benefit and recovery in locations where the 2012 drought deficit persisted. We evaluated these outcomes separately because regional storage integrates groundwater gains across the aquifer, whereas local recovery depends on both the magnitude of the water-level response and how far a location remains from its pre-drought state.

For the 2013 intervention, storage was evaluated at the end of the May–October pumping reduction, two months after historical pumping resumed and six months after pumping resumed (Fig. S23). This allowed us to distinguish the immediate effect of the spatial allocation from the part of that benefit that remained after the intervention. The Leverage-guided allocation retained a larger regional storage benefit than Uniform and High-pumping allocations at all three evaluation times, showing that its regional advantage was not limited to the intervention endpoint.

For the 2012 drought-period experiment, we also measured how much of the drought-related storage deficit was offset by reduced pumping. The historical December 2012 deficit was calculated relative to the pre-drought WTD defined in Section S1.5:

$$S_{def}^{hist} = \sum_i A_i S_{y,i} \left[D_i^{hist} \left(Dec\ 2012 \right) - D_i^{pre} \right]_+$$

where $[x]_+ = \max(x, 0)$. The fraction of this deficit offset under allocation k was

$$F_{offset,k} = 100 \frac{\Delta S_k \left(Dec\ 2012 \right)}{S_{def}^{hist}}$$

We compared this quantity among the three allocations and across conservation budgets (Fig. S24). The additional deficit offset from Leverage-guided allocation remained positive relative to both Uniform and High-pumping allocations at the 10%, 20% and 30% budgets, indicating that the regional advantage was not specific to the 20% budget used in the main comparison.

Intervention timing was evaluated separately using Uniform reductions applied over the same May–October window in 2012 and 2013 and matched by conserved pumping volume. We expressed the response as storage benefit per unit pumping reduction,

$$\eta_k(t) = \frac{\Delta S_k(t)}{\Delta V_k}$$

and evaluated it at the intervention endpoint and six months after historical pumping resumed (Fig. S26). The larger storage return during the 2013 recovery-period intervention remained evident six months later.

Local recovery was evaluated independently of regional storage. We used the half-recovery threshold defined from the historical drought trajectory in Section S1.5. Cells that had not reached this threshold before the May 2013 intervention were defined as initially unrecovered, and this group was kept fixed for all pumping allocations. For each cell, we compared the month in which the half-recovery threshold was reached under reduced and historical pumping. A positive recovery-time gain indicated that reduced pumping allowed the threshold to be reached earlier. We also tracked the cumulative fraction of initially unrecovered cells reaching half-recovery through time

(Fig. S25). Among cells whose recovery was advanced, most gains occurred over months to about one year, whereas multi-year advances were uncommon.

Threshold crossing alone does not describe how much an unrecovered location improves. We therefore also evaluated WTD improvement directly within the fixed initially unrecovered group. The initial recovery gap was defined as

$$G_i = D_i^{hist}(t_0) - D_i^{50}$$

where t_0 is the start of the May 2013 intervention and D_i^{50} is the historical half-recovery threshold. Larger G_i indicates that a location remained farther from half-recovery when the intervention began. We compared WTD improvement among four recovery-gap intervals and examined how positive benefit was distributed across initially unrecovered cells (Fig. S27). The Leverage-guided response decreased as the initial recovery gap increased and was smallest among locations more than 1 m from half-recovery. Positive benefit was also concentrated in a subset of initially unrecovered cells, showing that an aggregate improvement within this group did not imply an even distribution of recovery benefit.

Finally, we examined where the additional regional storage produced by Leverage-guided allocation occurred. For each cell, the difference relative to Uniform allocation was

$$\Delta s_i^{L-U}(t) = A_i S_{y,i} \left[\delta D_{i,L}(t) - \delta D_{i,U}(t)\right]$$

These differences were summed separately for locations that remained unrecovered at the start of the 2013 intervention and locations that had already reached half-recovery. This decomposition underlies Fig. 4e and distinguishes an increase in total regional storage from an increase in storage within locations where the drought deficit was still present.

Figs. S1 to S27

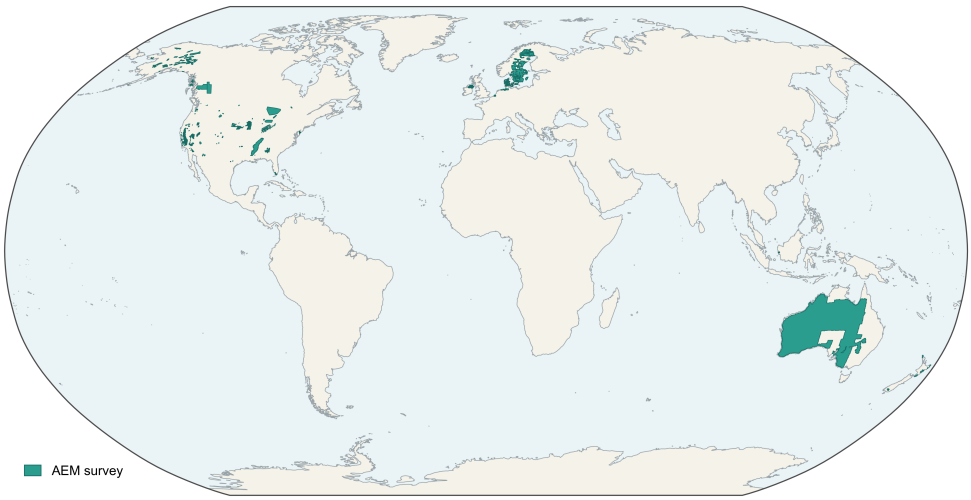

Fig. S1 Global distribution of public airborne electromagnetic survey datasets across the United States [4,42–44], Sweden [45,46], the Netherlands [47–49], Northern Ireland (UK) [50,51], Australia [52,53], Canada [54,55], New Zealand [56,57], Denmark [58,59] and Indonesia [60,61].

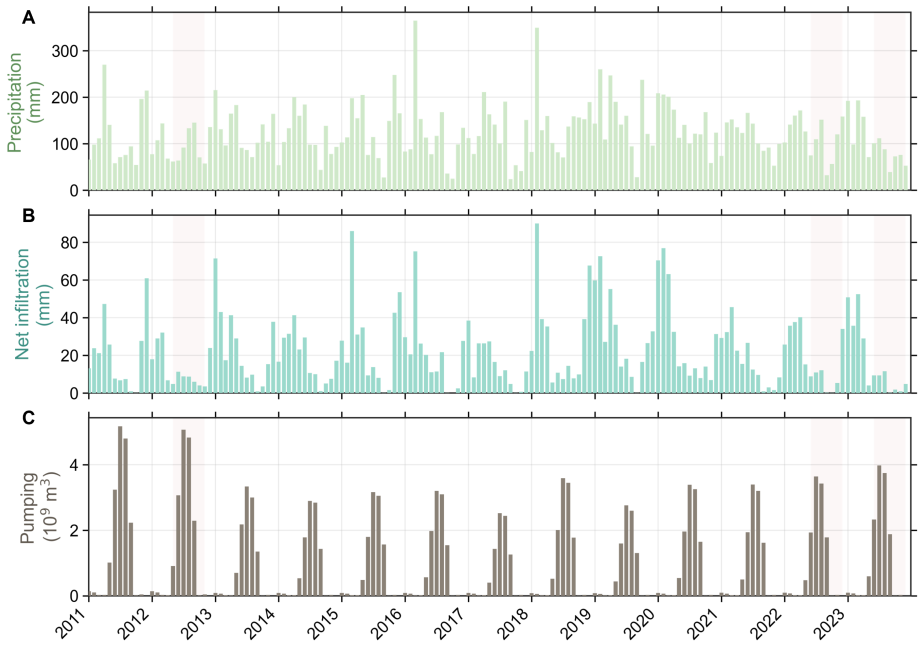

Fig. S2 Regional hydroclimatic data and groundwater pumping data from 2011 to 2023. a, Monthly precipitation. b, Monthly net infiltration. c, Monthly groundwater pumping across the MRVA.

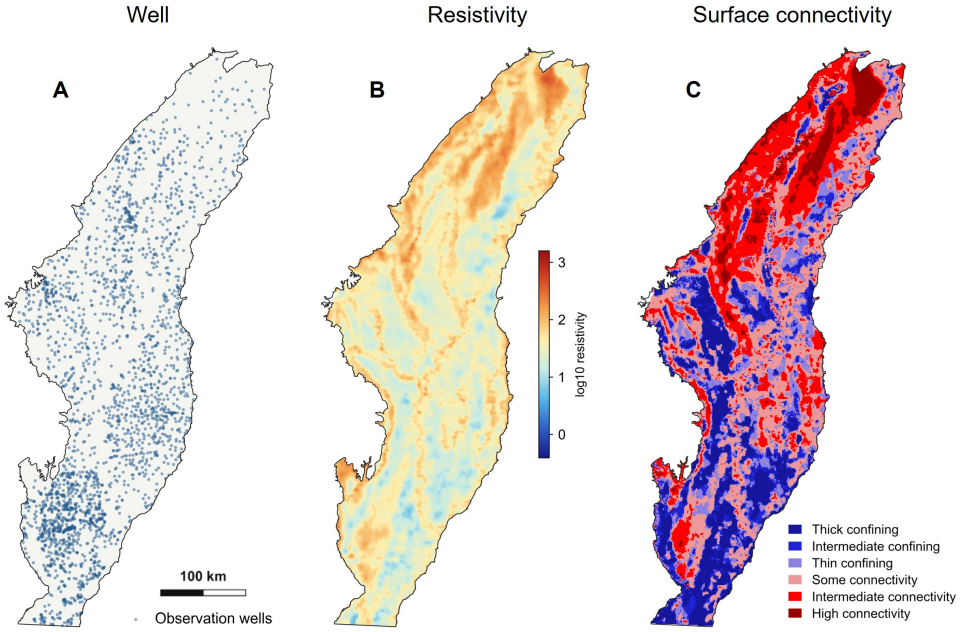

Fig. S3 Spatial inputs for the MRVA reconstruction. a, Observation wells. b, Surface resistivity. c, Surface connectivity.

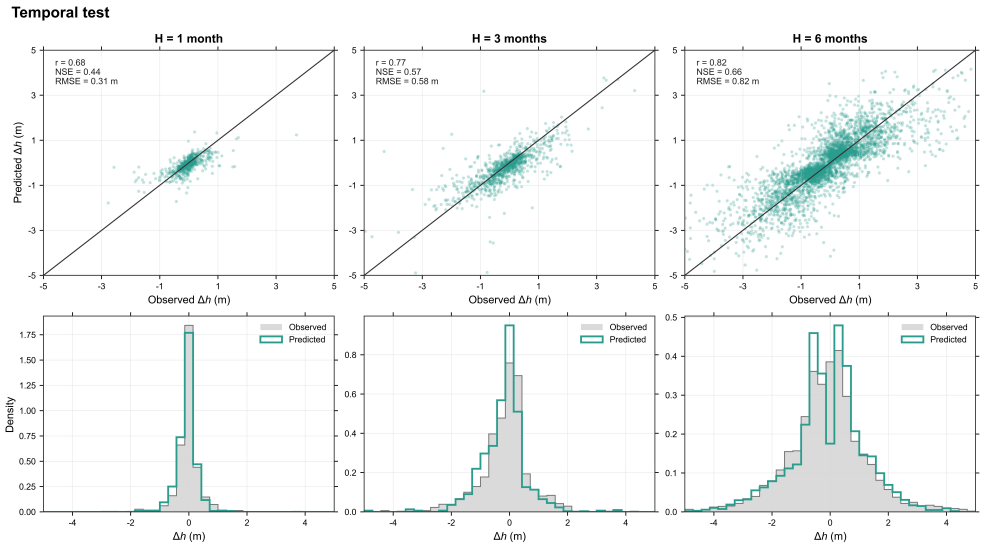

Fig. S4 GNN performance in the temporal test set. The upper row shows observed and predicted WTD changes for H = 1, 3 and 6 months. The lower row shows the corresponding distributions of observed and predicted changes.

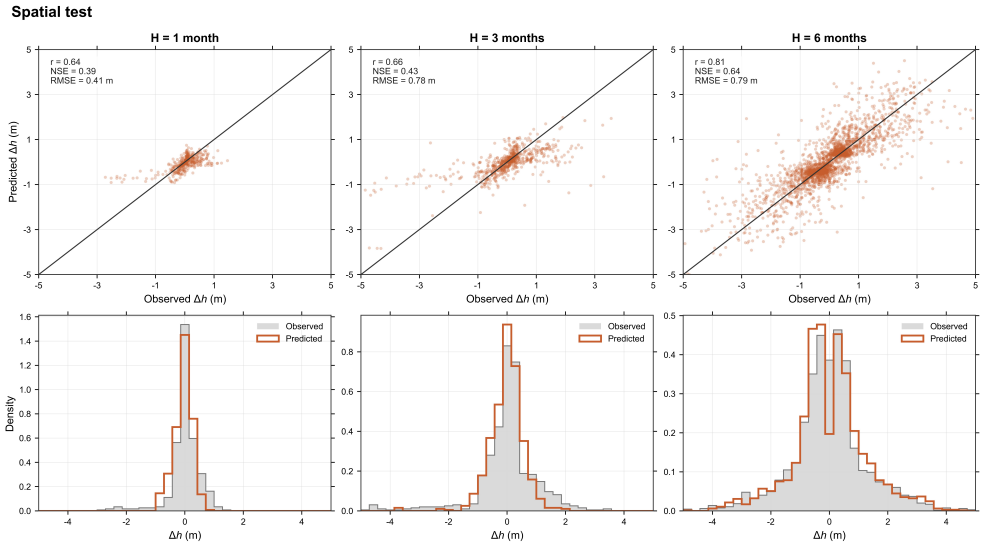

Fig. S5 GNN performance in the spatial test set. The upper row shows observed and predicted WTD changes for H = 1, 3 and 6 months. The lower row shows the corresponding distributions of observed and predicted changes.

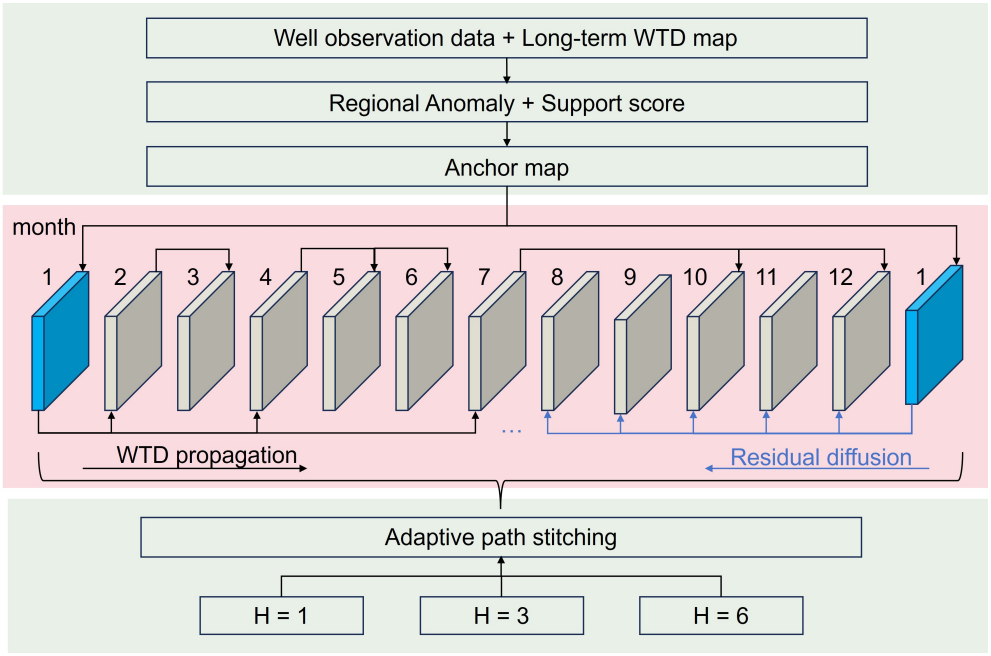

Fig. S6 Monthly WTD reconstruction workflow. Well observations and the long-term WTD map were combined to construct anchor maps. The reconstruction between anchors used 1-, 3- and 6-month GNN predictions, adaptive path stitching and residual diffusion. Well observations were then used for local constraints and the regional groundwater-storage adjustment.

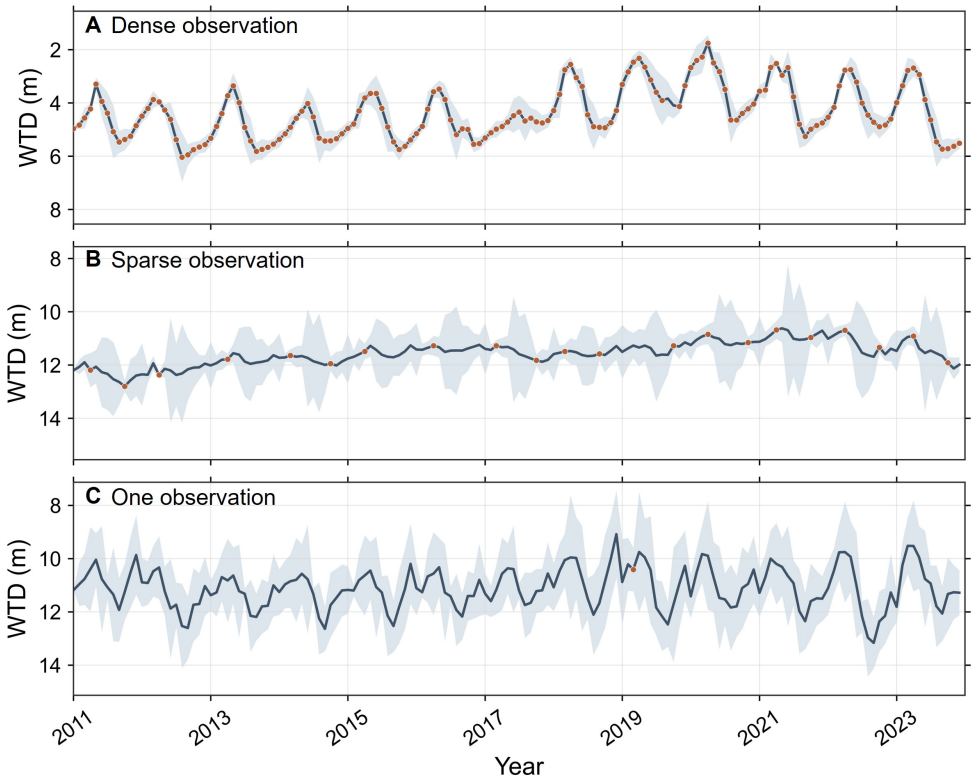

Fig. S7 Reconstructed WTD at locations with different observational support. The panels show monthly WTD for a location with dense observations, a location with sparse observations and a location with one observation during 2011–2023. Shaded bands show the PI75 intervals, and points show the available observations.

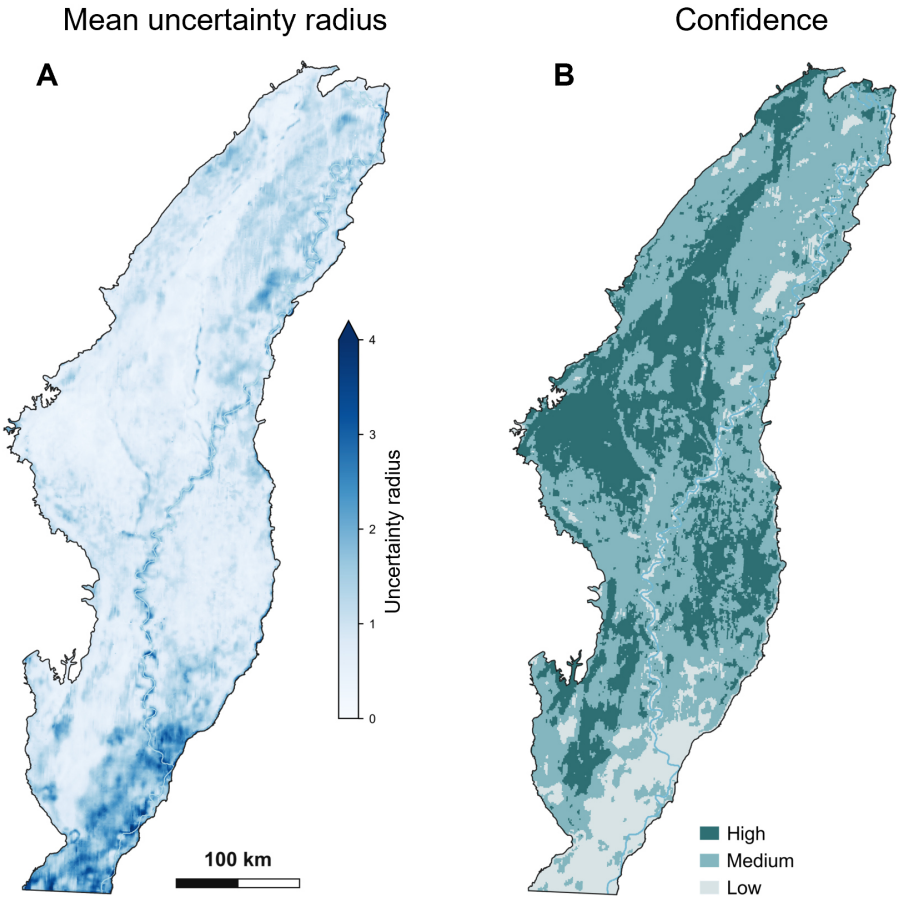

Fig. S8 Spatial distribution of WTD reconstruction uncertainty. a, Mean PI75 radius across the 156 reconstructed months. b, WTD reconstruction confidence based on the mean PI75 radius. High confidence denotes a mean PI75 radius of 0.70 m or less, medium confidence a radius greater than 0.70 m and no greater than 1.50 m, and low confidence a radius greater than 1.50 m.

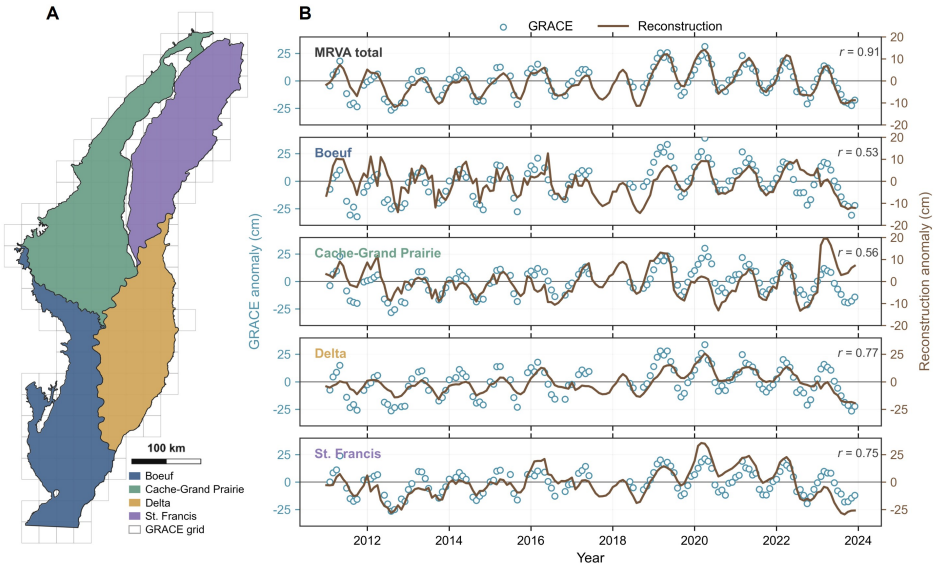

Fig. S9 Comparison of reconstructed and GRACE terrestrial water storage anomaly. The left panel shows the MRVA regions and GRACE grid cells. The right panels show the reconstructed and GRACE anomalies for the MRVA and four subregions.

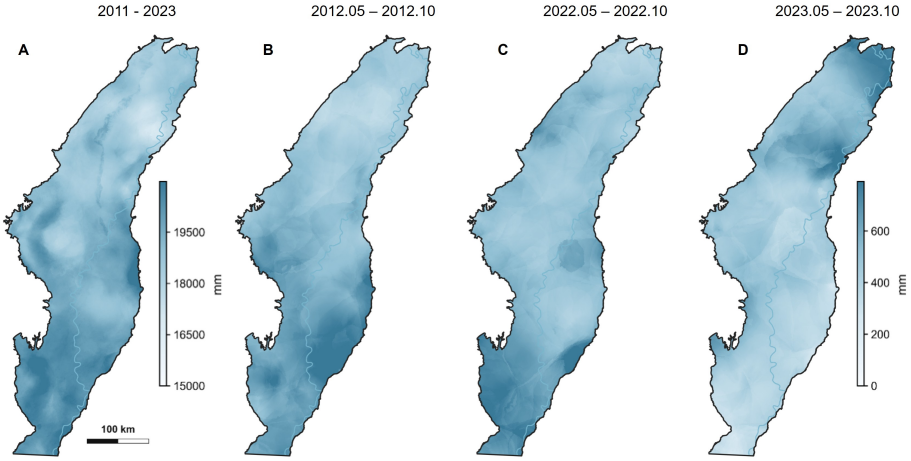

Fig. S10 Precipitation during the reference and drought periods. a, 2011–2023. b, May–October 2012. c, May–October 2022. d, May–October 2023.

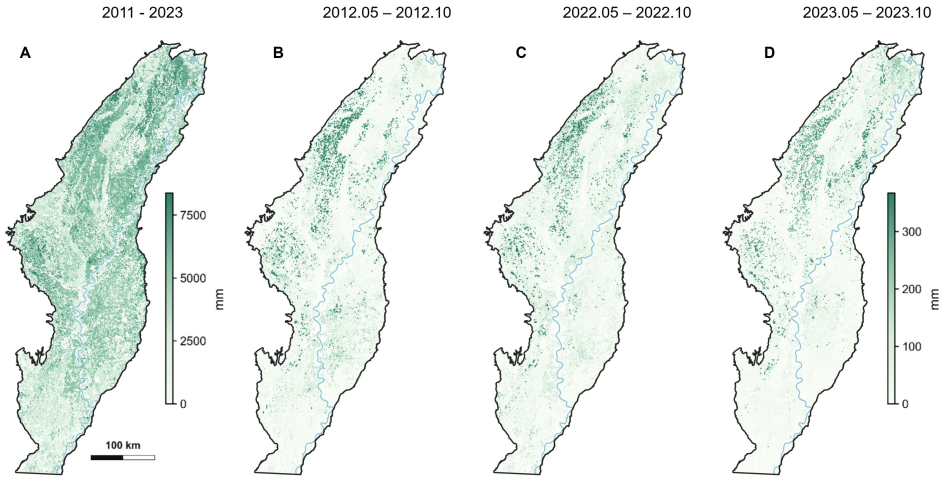

Fig. S11 Net infiltration during the reference and drought periods. a, 2011–2023. b, May–October 2012. c, May–October 2022. d, May–October 2023.

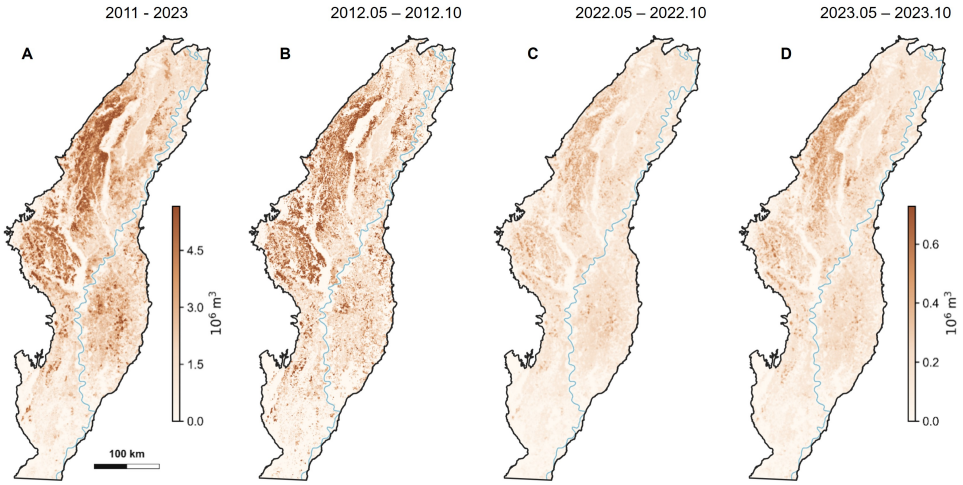

Fig. S12 Groundwater pumping during the reference and drought periods. a, 2011–2023. b, May–October 2012. c, May–October 2022. d, May–October 2023.

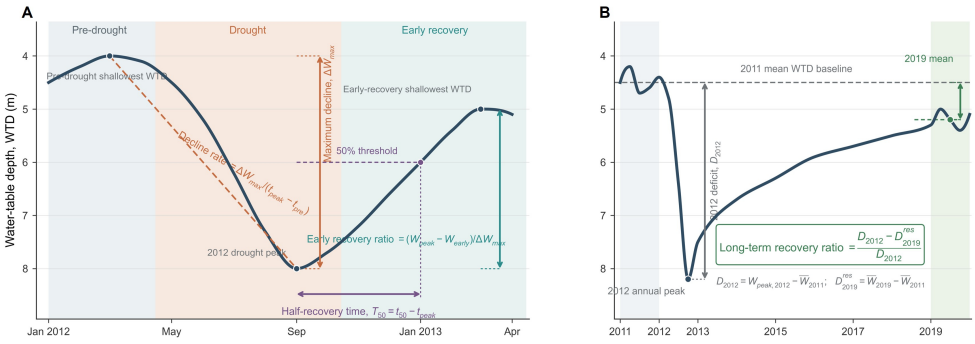

Fig. S13 Definitions of the five drought-response metrics. a, Maximum decline, decline rate, early recovery ratio and half-recovery time during the 2012 drought and early-recovery window. b, Long-term recovery ratio calculated from the 2011 mean WTD, the 2012 annual peak and the 2019 mean WTD.

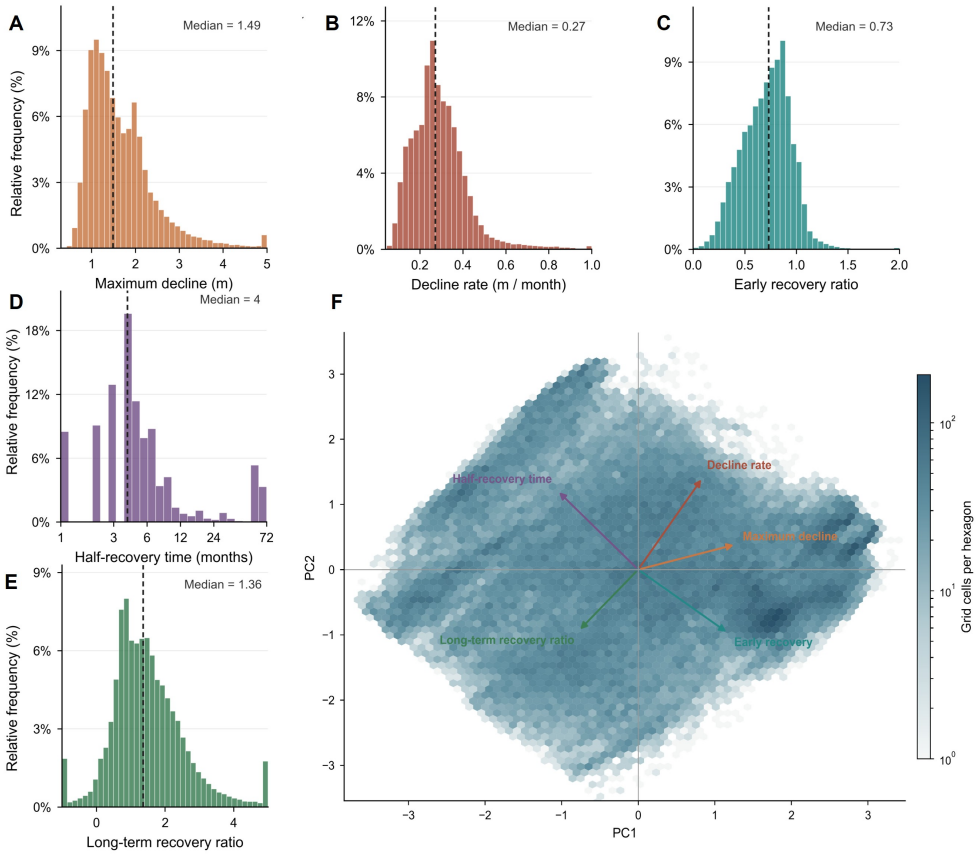

Fig. S14 Distributions and response space of the five drought-response metrics. a–e, Distributions of maximum decline, decline rate, early recovery ratio, half-recovery time and long-term recovery ratio. f, Principal-component response space based on the five metrics. The response-space plot uses complete metric records before response classification and does not show class boundaries.

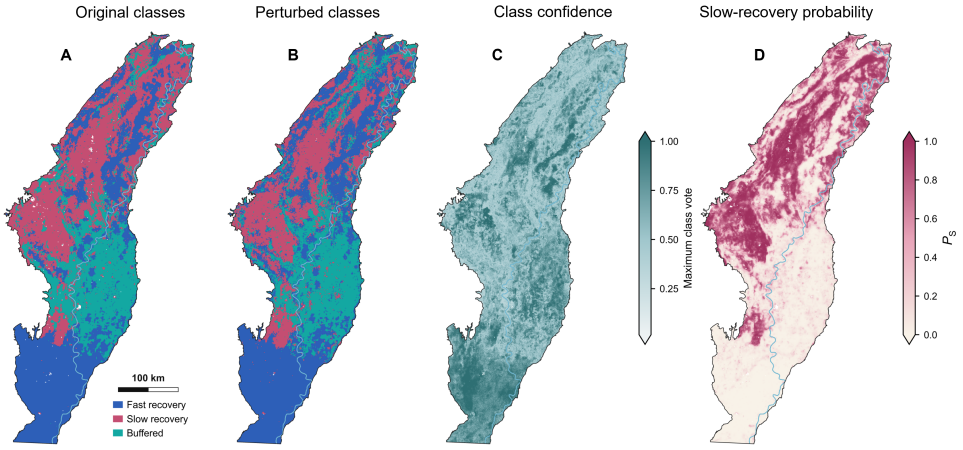

Fig. S15 Stability of the groundwater-response zones and mapped Slow-recovery probability. a, Recovery zones derived from the original WTD reconstruction. b, Most frequent zone across 64 perturbed reconstructions. c, Classification confidence, defined as the fraction of perturbations assigned to the most frequent zone. d, Slow-recovery probability predicted from AEM resistivity and pumping.

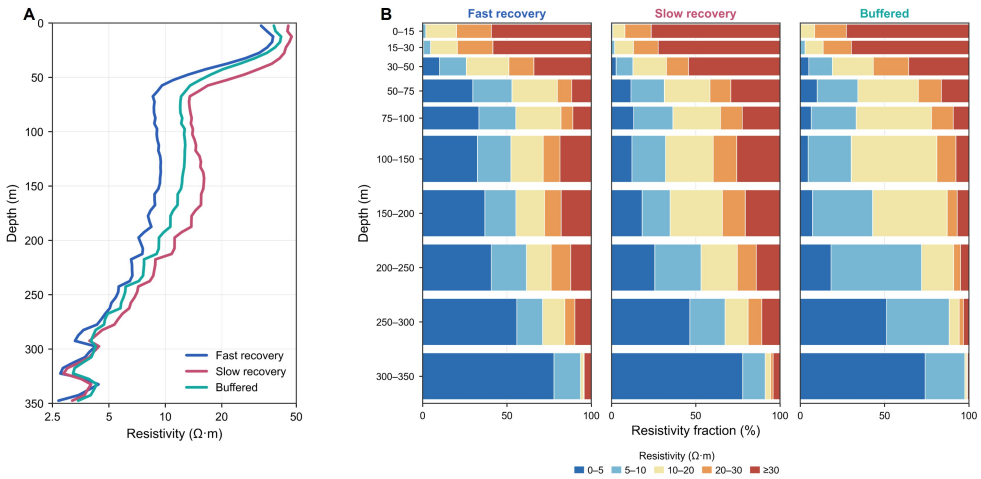

Fig. S16 Resistivity structure of the three response zones. a, Median resistivity profiles from 0 to 350 m below land surface. b, Proportions of resistivity ranges within ten depth intervals. The resistivity ranges are 0–5, 5–10, 10–20, 20–30 and $\geq$30 Ω·m.

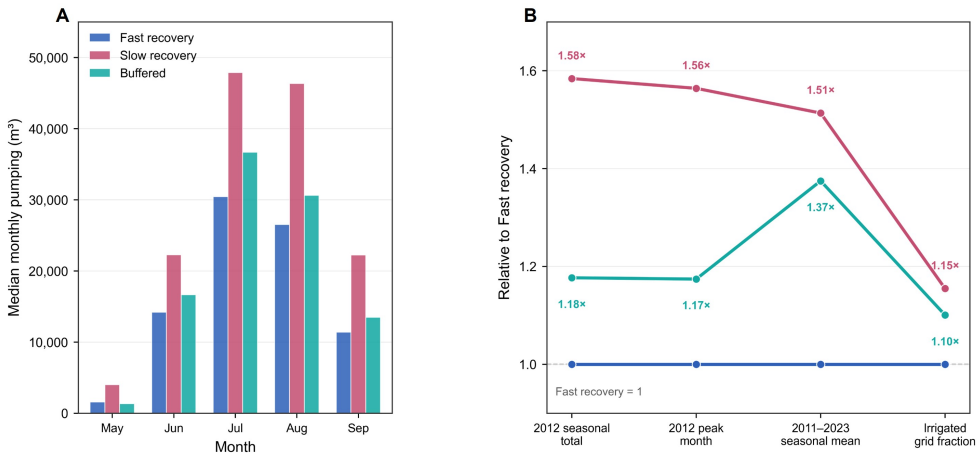

Fig. S17 Pumping differences among the three response classes. a, Median monthly pumping during May–September 2012. b, Pumping relative to the fast-recovery class for total May–October 2012 pumping, maximum monthly May–September 2012 pumping, mean May–September pumping during 2011–2023 and the proportion of grid cells with pumping during May–September 2012.

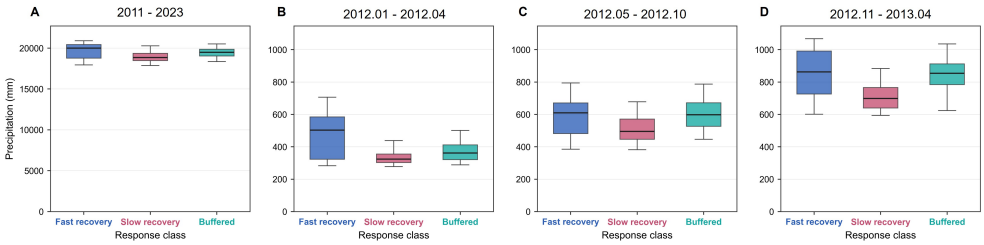

Fig. S18 Precipitation differences among the three response classes. a, 2011–2023 total precipitation. b, January–April 2012. c, May–October 2012. d, November 2012–April 2013.

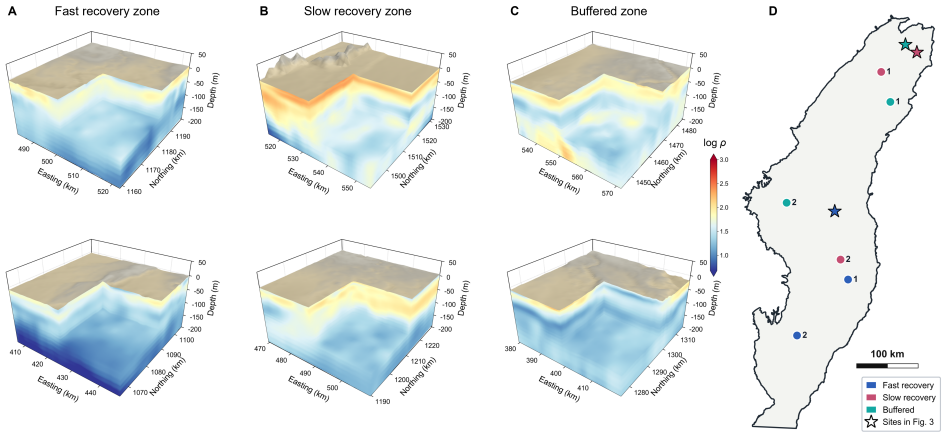

Fig. S19 Representative resistivity structures and locations used for class interpretation. a–c, Three-dimensional resistivity sections for Fast recovery, Slow recovery and Buffered examples. The upper and lower rows show two locations for each response class. d, Locations of the representative sites within the MRVA.

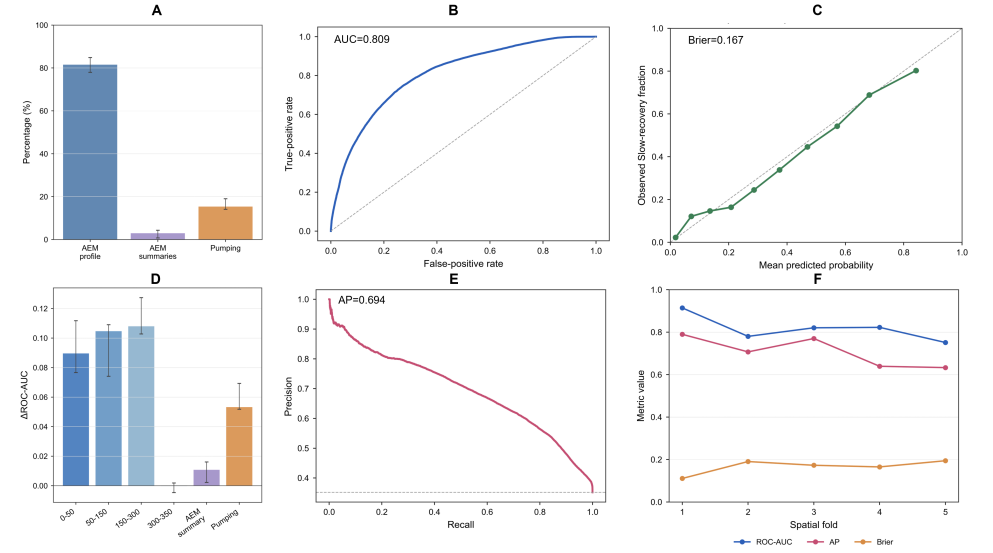

Fig. S20 Performance and predictor information for the Slow-recovery probability model. a, Grouped permutation importance of the AEM resistivity profile, AEM profile summaries and pumping. b, Receiver operating characteristic curve from spatial-block validation. c, Reliability of the predicted Slow-recovery probabilities. d, Decrease in ROC-AUC after jointly permuting AEM depth intervals and pumping. e, Precision–recall curve. f, ROC-AUC, average precision and Brier score across the five spatial folds.

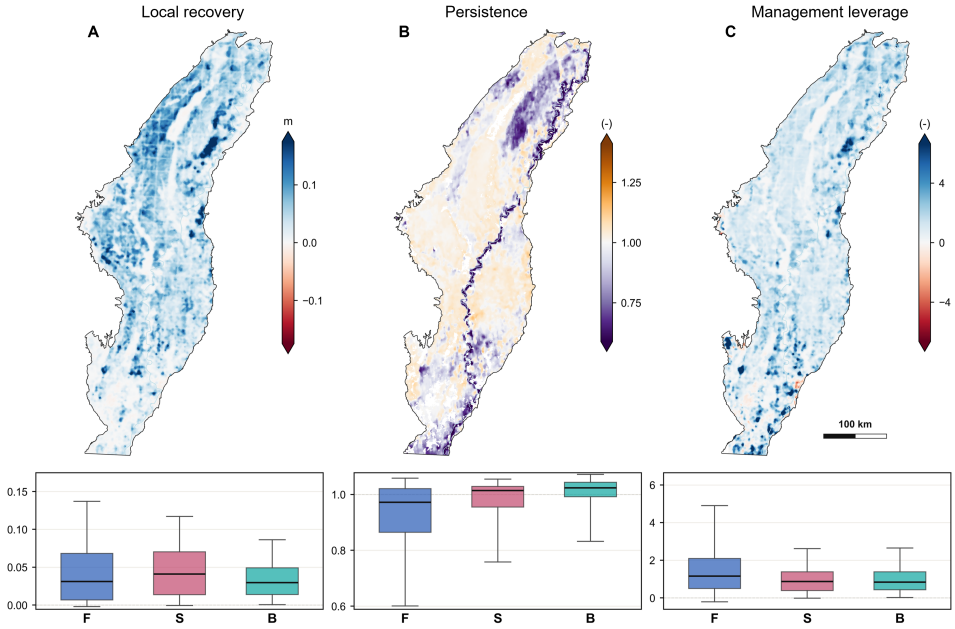

Fig. S21 Spatial variation in groundwater response to reduced pumping. a, Local recovery in December 2012 after pumping was reduced by 20% in each management unit separately during 2012. b, Persistence of the December 2012 response one year after historical pumping resumed. c, Management leverage, defined as local recovery divided by the equivalent pumping-reduction depth. Box plots show the corresponding distributions across Fast-recovery, Slow-recovery and Buffered zones.

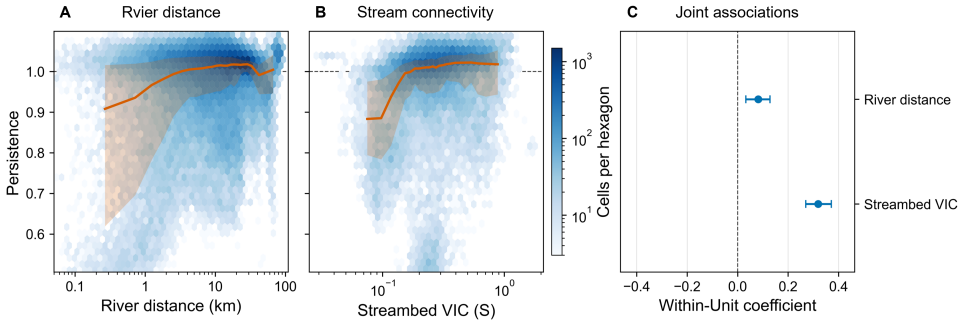

Fig. S22 Persistence in relation to river distance and streambed connectivity. a, Persistence as a function of distance to major rivers. b, Persistence as a function of streambed vertical electrical conductance (VIC); lower VIC indicates greater potential connection between the stream and aquifer. Hexagons show the density of observations and orange curves summarize the central trend. c, Standardized coefficients from the joint within-management-unit rank analysis of river distance and streambed VIC.

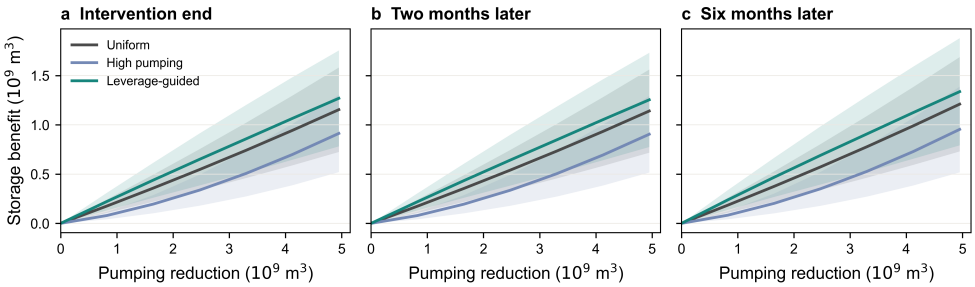

Fig. S23 Storage benefit under alternative pumping allocations during the 2013 recovery period. a–c, Storage benefit at the end of the May–October 2013 intervention, two months after historical pumping resumed and six months after historical pumping resumed, respectively. Curves compare Uniform, High-pumping and Leverage-guided allocations at the same actual pumping-reduction volume. Lines show the mean across the five fitted models and shading shows variation among them.

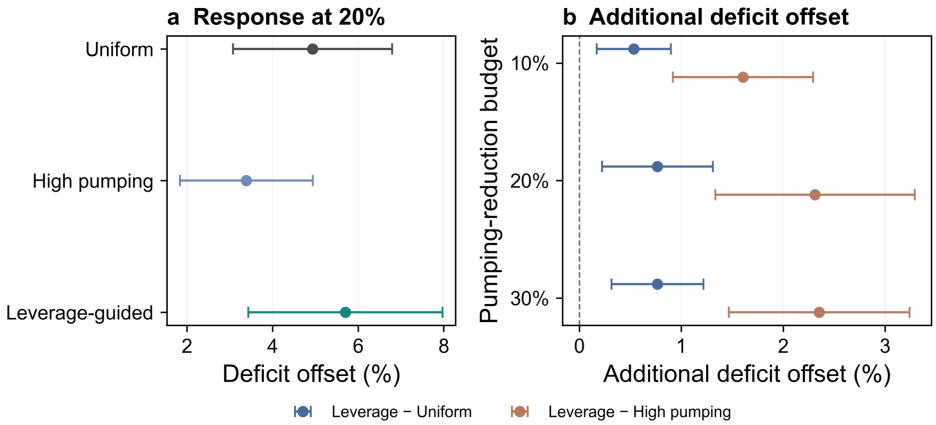

Fig. S24 Reduction of the December 2012 groundwater deficit under alternative pumping allocations. a, Fraction of the historical December 2012 storage deficit offset at the common 20% conservation budget. b, Additional deficit offset under Leverage-guided allocation relative to Uniform and High-pumping allocations at 10%, 20% and 30% conservation budgets. Points show the mean across the five fitted models; horizontal bars show their spread.

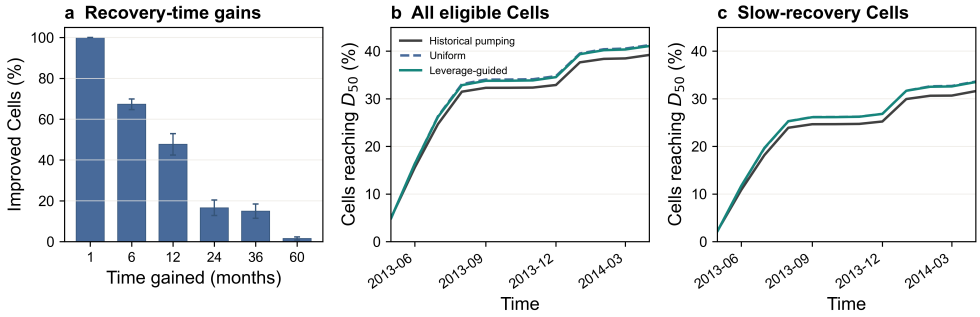

Fig. S25 Changes in half-recovery timing during the 2013 recovery-period intervention. a, Recovery-time gains under the 20% Uniform allocation, expressed as the percentage of initially unrecovered cells reaching the historical half-recovery threshold (D_{50}) earlier than under historical pumping by at least the indicated number of months. b, Cumulative fraction of initially unrecovered cells reaching D_{50} under historical pumping, Uniform and Leverage-guided allocations. c, Same comparison for Slow-recovery cells. Initially unrecovered cells had not reached D_{50} before the May 2013 intervention. Lines show model means; error bars in a show variation among the five fitted models.

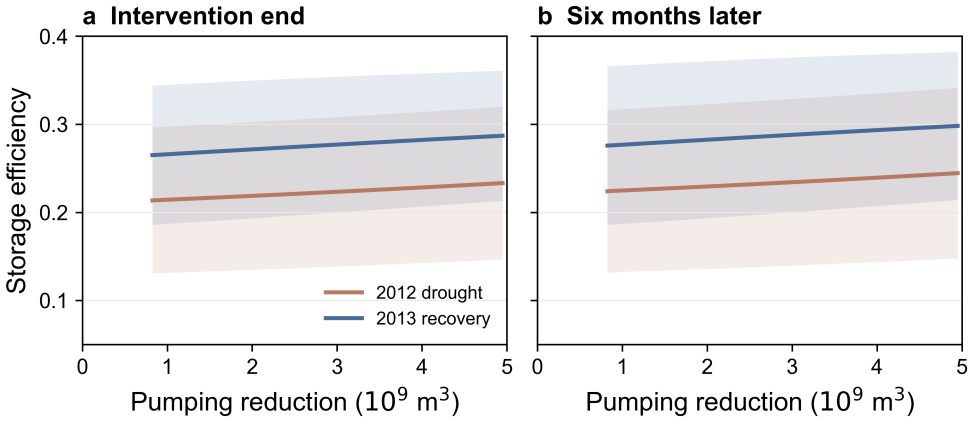

Fig. S26 Hydrologic return of pumping reductions during the drought and recovery periods. a, Storage benefit per unit pumping reduction at the end of the May–October intervention for Uniform reductions applied in 2012 and 2013. b, Same comparison six months after historical pumping resumed. Lines show the mean across the five fitted models and shading shows variation among them.

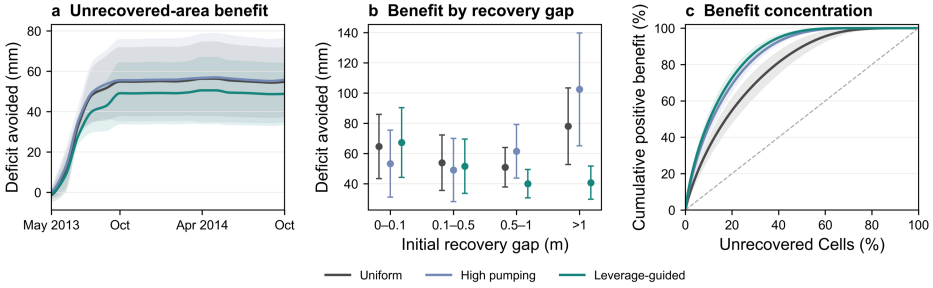

Fig. S27 Groundwater benefits in locations that had not reached half-recovery at the start of the 2013 intervention. a, Mean water-table deficit avoided during and after the May–October 2013 pumping reduction under Uniform, High-pumping and Leverage-guided allocations. b, Deficit avoided at the end of the intervention across four initial recovery-gap intervals, defined by the distance from the historical half-recovery threshold. c, Cumulative share of positive groundwater benefit across initially unrecovered cells, ranked from largest to smallest individual benefit. The dashed line in c represents an equal distribution of benefit among cells. Lines and points show model means; shading and error bars indicate variation among the five fitted models.

Tables S1 to S3

Table S1 | Performance of the GNN models across prediction intervals

H	Temporal test RMSE			Spatial test RMSE		
	(m)	r	NSE	(m)	r	NSE
1	0.314	0.676	0.444	0.411	0.638	0.391
3	0.584	0.773	0.575	0.783	0.663	0.430
6	0.820	0.819	0.663	0.788	0.812	0.638

Table S2 | Target percentile ranks and metric weights used to define the response prototypes. Values are target percentile ranks; values in parentheses are metric weights.

Response group	Maximum decline	Decline rate	Early recovery ratio	Half-recovery time	Long-term deficit
Fast recovery	0.78 (0.20)	0.65 (0.16)	0.88 (0.32)	0.12 (0.24)	0.35 (0.08)
Slow recovery	0.62 (0.15)	0.55 (0.10)	0.25 (0.25)	0.88 (0.35)	0.80 (0.15)
Buffered	0.10 (0.35)	0.12 (0.30)	0.60 (0.08)	0.28 (0.07)	0.12 (0.20)

Table S3 | Drought-response metrics across the three response classes

Response class	Metric	Median	Q25-Q75	Q5-Q95
	Maximum decline (m)	2.025	1.738–2.437	1.362–3.763
	Decline rate (m/month)	0.305	0.254–0.388	0.218–0.640
Fast recovery	Early recovery ratio	0.882	0.816–0.962	0.723–1.081
	Half-recovery time (month)	3.000	2.000–4.000	1.000–5.000
	Long-term recovery ratio	1.052	0.765–1.475	0.264–2.164
	Maximum decline (m)	1.347	1.110–1.699	0.885–2.555
	Decline rate (m/month)	0.325	0.267–0.378	0.184–0.466
Slow recovery	Early recovery ratio	0.508	0.358–0.644	0.173–0.767
	Half-recovery time (month)	8.000	6.000–58.000	4.000–67.000
	Long-term recovery ratio	1.193	0.611–1.689	−0.837–3.392
	Maximum decline (m)	1.035	0.885–1.204	0.693–1.455
	Decline rate (m/month)	0.165	0.131–0.207	0.093–0.282
Buffered	Early recovery ratio	0.660	0.527–0.791	0.348–1.097
	Half-recovery time (month)	5.000	3.000–7.000	2.000–32.000
	Long-term recovery ratio	2.272	1.827–2.781	0.582–4.408

References

1. Painter, J. A. & Westerman, D. A. Mississippi Alluvial Plain Extent, November 2017. U.S. Geological Survey data release (2018). https://doi.org/10.5066/F70R9NMJ.
2. U.S. Geological Survey. USGS Water Data for the Nation: U.S. Geological Survey National Water Information System database. https://doi.org/10.5066/F7P55KJN (accessed 29 April 2026)
3. U.S. Geological Survey. USGS Water Data APIs. https://api.waterdata.usgs.gov/ (accessed 20 October 2026)
4. James, S. R. & Minsley, B. J. Combined results and derivative products of hydrogeologic structure and properties from airborne electromagnetic surveys in the

Mississippi Alluvial Plain (ver. 2.0, July 2024). U.S. Geological Survey data release (2021). https://doi.org/10.5066/P9382RCI

5. Burton, B. L. et al. Airborne electromagnetic, magnetic, and radiometric surveys of the Mississippi Alluvial Plain and Chicot Aquifer System, March 2018 and May–August 2021. U.S. Geological Survey data release (2024). https://doi.org/10.5066/P9KPK3UJ.
6. Hoogenboom, B. E., Minsley, B. J., James, S. R. & Pace, M. D. Airborne electromagnetic, magnetic, and radiometric survey of the Mississippi Alluvial Plain, Mississippi Embayment, and Gulf Coastal Plain, September 2021–January 2022. U.S. Geological Survey data release (2023). https://doi.org/10.5066/P93DO0EO.
7. Burton, B. L., Minsley, B. J., Bloss, B. R. & Kress, W. H. Airborne electromagnetic, magnetic, and radiometric survey of the Mississippi Alluvial Plain, November 2018–February 2019. U.S. Geological Survey data release (2021). https://doi.org/10.5066/P9XBBBUU.
8. Minsley, B. J. et al. Airborne electromagnetic, magnetic, and radiometric survey of the Mississippi Alluvial Plain, November 2019–March 2020. U.S. Geological Survey data release (2021). https://doi.org/10.5066/P9E44CTQ.
9. Clark, B. R. et al. National-scale grid to support regional groundwater availability studies and a national hydrogeologic framework. U.S. Geological Survey data release (2018). https://doi.org/10.5066/F7P84B24
10. U.S. Geological Survey. 1/3rd arc-second Digital Elevation Models (DEMs) - USGS National Map 3DEP Downloadable Data Collection. U.S. Geological Survey data release. https://data.usgs.gov/datacatalog/data/USGS:3a81321b-c153-416f-98b7-cc8e5f0e17c3 (accessed 29 April 2026)
11. Villers, J. J., Haugh, C. M. & Duncan, L. L. Updated Estimates of Water Budget Components for the Mississippi Embayment Region Using a Soil-Water-Balance

Model, 2000–2024. U.S. Geological Survey data release (2026). https://doi.org/10.5066/P1YV4WD7.

12. Nielsen, M. G. & Westenbroek, S. M. Updated estimates of water budget components for the Mississippi Embayment Region using a Soil-Water-Balance model, 2000–2020. U.S. Geological Survey Scientific Investigations Report 2023-5080 (2023). https://doi.org/10.3133/sir20235080
13. Haugh, C. M. Aquaculture and Irrigation Water-Use Model (AIWUM) version 2.0 updated input and output datasets for the Mississippi Alluvial Plain, 2014–2023. U.S. Geological Survey data release (2026). https://doi.org/10.5066/P13PQXTH
14. Bristow, E. L. & Wilson, J. L. Aquaculture and irrigation water-use model (AIWUM) version 1.1 estimates and related datasets for the Mississippi Alluvial Plain. U.S. Geological Survey data release (2023). https://doi.org/10.5066/P9RGZOBZ
15. Thornton, M. M., Shrestha, R., Wei, Y., Thornton, P. E. & Kao, S.-C. Daymet: Monthly Climate Summaries on a 1-km Grid for North America, Version 4 R1. ORNL DAAC (2022). https://doi.org/10.3334/ORNLDAAC/2131.
16. Ma, Y. et al. High resolution US water table depth estimates reveal quantity of accessible groundwater. Commun. Earth Environ. 7, 45 (2026). https://doi.org/10.1038/s43247-025-03094-3
17. Breiman, L. Random forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
18. Tapley, B. D., Bettadpur, S., Watkins, M. & Reigber, C. The gravity recovery and climate experiment: mission overview and early results. Geophys. Res. Lett. 31, L09607 (2004). https://doi.org/10.1029/2004GL019920
19. Save, H. CSR GRACE and GRACE-FO RL06 Mascon Solutions v02. Center for Space Research, University of Texas at Austin (2020). https://doi.org/10.15781/cgq9-nh24

20. Lehner, B. & Grill, G. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrol. Process. 27, 2171–2186 (2013). https://doi.org/10.1002/hyp.9740

21. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In Proceedings of the 5th International Conference on Learning Representations (2017).

22. Roberts, D. R. et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929 (2017). https://doi.org/10.1111/ecog.02881

23. Nash, J. E. & Sutcliffe, J. V. River flow forecasting through conceptual models part I—A discussion of principles. Journal of Hydrology 10, 282–290 (1970). https://doi.org/10.1016/0022-1694(70)90255-6

24. Healy, R. W. & Cook, P. G. Using groundwater levels to estimate recharge. Hydrogeology Journal 10, 91–109 (2002). https://doi.org/10.1007/s10040-001-0178-0

25. Shafer, G. & Vovk, V. A tutorial on conformal prediction. Journal of Machine Learning Research 9, 371–421 (2008). https://www.jmlr.org/papers/v9/shafer08a.html

26. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems 30, 6402–6413 (2017). https://papers.nips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html

27. Van Loon, A. F. et al. Drought in the Anthropocene. Nat. Geosci. 9, 89–91 (2016). https://doi.org/10.1038/ngeo2646

28. Cuthbert, M. O. et al. Observed controls on resilience of groundwater to climate variability in sub-Saharan Africa. Nature 572, 230–234 (2019). https://doi.org/10.1038/s41586-019-1441-7

29. Bloomfield, J. P. & Marchant, B. P. Analysis of groundwater drought building on the standardised precipitation index approach. Hydrology and Earth System Sciences 17, 4769–4787 (2013). https://doi.org/10.5194/hess-17-4769-2013

30. National Drought Mitigation Center, U.S. Department of Agriculture, National Oceanic and Atmospheric Administration & National Aeronautics and Space Administration. U.S. Drought Monitor data archive. https://droughtmonitor.unl.edu/DmData/DataDownload.aspx (accessed 22 July 2026).

31. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A 374, 20150202 (2016). https://doi.org/10.1098/rsta.2015.0202

32. Ikard, S. J., Minsley, B. J., Rigby, J. R. & Kress, W. H. A model of transmissivity and hydraulic conductivity from electrical resistivity distribution derived from airborne electromagnetic surveys of the Mississippi River Valley Alluvial Aquifer, Midwest USA. Hydrogeol. J. 31, 313–334 (2023). https://doi.org/10.1007/s10040-022-02590-6

33. Palacky, G. J. Resistivity characteristics of geologic targets. In Electromagnetic Methods in Applied Geophysics, Vol. 1 (ed. Nabighian, M. N.) chap. 3 (Society of Exploration Geophysicists, 1988). https://doi.org/10.1190/1.9781560802631.ch3

34. Archie, G. E. The electrical resistivity log as an aid in determining some reservoir characteristics. Transactions of the AIME 146, 54–62 (1942). https://doi.org/10.2118/942054-G

35. Waxman, M. H. & Smits, L. J. M. Electrical conductivities in oil-bearing shaly sands. Society of Petroleum Engineers Journal 8, 107–122 (1968). https://doi.org/10.2118/1863-A

36. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. Mach. Learn. 63, 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1

37. Fawcett, T. An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874 (2006). https://doi.org/10.1016/j.patrec.2005.10.010

38. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10, e0118432 (2015). https://doi.org/10.1371/journal.pone.0118432

39. Brier, G. W. Verification of forecasts expressed in terms of probability. Monthly Weather Review 78, 1–3 (1950). https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2

40. Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning, 625–632 (2005). https://doi.org/10.1145/1102351.1102430

41. Fisher, A., Rudin, C. & Dominici, F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research 20, 1–81 (2019). https://www.jmlr.org/papers/v20/18-760.html

42. Minsley, B. J. et al. Airborne geophysical surveys of the lower Mississippi Valley demonstrate system-scale mapping of subsurface architecture. Commun. Earth Environ. 2, 131 (2021). https://doi.org/10.1038/s43247-021-00200-z

43. Johnson, M. R. et al. Airborne Geophysical Survey Inventory of the conterminous United States, Alaska, Hawaii, and Puerto Rico (ver. 5.0, April 2024). U.S. Geological Survey data release (2019). https://doi.org/10.5066/P9K8YTW1

44. Dlubac, K., Springhorn, S., Brezing, B., Altare, C. & Godwin, T. California's statewide airborne electromagnetic surveys and preliminary hydrogeologic interpretations. California Geology 55, 18–31 (2024). https://www.conservation.ca.gov/cgs/documents/publications/california-geology-magazine/CG_2024-AEMsurveys.pdf

45. Pedersen, L. B., Persson, L., Bastani, M. & Byström, S. Airborne VLF measurements and mapping of ground conductivity in Sweden. J. Appl. Geophys. 67, 250–258 (2009). https://doi.org/10.1016/j.jappgeo.2007.11.002.

46. Geological Survey of Sweden. Electrical conductivity. SGU https://www.sgu.se/en/about-sgu/tasks-and-activities-new/geological-mapping-surveys/geophysical-mapping/electrical-conductivity/ (accessed 22 July 2026).

47. Delsman, J. R. et al. Large-scale, probabilistic salinity mapping using airborne electromagnetics for groundwater management in Zeeland, the Netherlands. Environ. Res. Lett. 13, 084011 (2018). https://doi.org/10.1088/1748-9326/aad19e

48. van Baaren, E. S. et al. FRESHEM Zeeland—FREsh Salt groundwater distribution by Helicopter ElectroMagnetic survey in the Province of Zeeland. Deltares Report 1209220-000-BGS-0030 (2018).

49. Provincie Zeeland. FRESHEM Zeeland data and documentation. GitHub (accessed 22 July 2026). https://github.com/ProvincieZeeland/FRESHEM.

50. Beamish, D., Cuss, R. J., Lahti, M., Scheib, C. & Tartaras, E. The Tellus airborne geophysical survey of Northern Ireland: final processing report. British Geological Survey Internal Report IR/06/136 (2006). https://nora.nerc.ac.uk/7427/.

51. Geological Survey of Northern Ireland. The Tellus airborne surveys of Northern Ireland (2007). https://gsni-data.bgs.ac.uk/geonetwork/srv/api/records/43e0d940-c82e-40bb-abc0-a775158500b5 (accessed 22 July 2026).

52. Ley-Cooper, A. Y., Brodie, R. C. & Richardson, M. AusAEM: Australia's airborne electromagnetic continental-scale acquisition program. Explor. Geophys. 51, 193–202 (2020). https://doi.org/10.1080/08123985.2019.1694393.

53. Geoscience Australia. AusAEM. Exploring for the Future, Geoscience Australia. https://www.eftf.ga.gov.au/ausaem (accessed 20 October 2026).

54. Geotech Ltd. Report on a helicopter-borne versatile time domain electromagnetic (VTEM) geophysical survey: QUEST Project, central British Columbia. Geoscience

BC Report 2008-04 (Geoscience BC, 2008). https://www.geosciencebc.com/major-projects/quest/.

55. Aeroquest Ltd. Report on a helicopter-borne AeroTEM system electromagnetic and magnetic survey: QUEST West, central British Columbia, Canada. Geoscience BC Report 2009-06 (Geoscience BC, 2009). https://www.geosciencebc.com/major-projects/quest-west/.
56. Herpe, M. et al. An overview of SkyTEM surveys in New Zealand: data acquisition, community engagement, and results from Northland. Australian Society of Exploration Geophysicists Extended Abstracts 2023 (2), 8th International Airborne Electromagnetics Workshop, Fitzroy Island (2023). https://doi.org/10.5281/zenodo.10067951
57. Earth Sciences New Zealand–GNS Science. SkyTEM Aerial Aquifer Mapping. https://www.gns.cri.nz/research-projects/skytem/ (accessed 20 October 2026)
58. Barfod, A. A. S., Møller, I. & Christiansen, A. V. Compiling a national resistivity atlas of Denmark based on airborne and ground-based transient electromagnetic data. J. Appl. Geophys. 134, 199–209 (2016). https://doi.org/10.1016/j.jappgeo.2016.09.017
59. Geological Survey of Denmark and Greenland. National Geophysical Database—GERDA https://eng.geus.dk/products-services-facilities/data-and-maps/national-geophysical-database-gerda (accessed 20 October 2026).
60. Silvestri, S. et al. Quantification of peat thickness and stored carbon at the landscape scale in tropical peatlands: a comparison of airborne geophysics and an empirical topographic method. J. Geophys. Res. Earth Surf. 124, 3107–3123 (2019). https://doi.org/10.1029/2019JF005273.
61. Silvestri, S. et al. Airborne electromagnetic (AEM) raw data West Kalimantan Indonesia. Zenodo (2019). https://doi.org/10.5281/zenodo.3571918.