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.

Reliability Auditing and Capacity-Constrained Review for Wheat-Yield Decision Support under Temporal and Spatial Shift

Xin Li

lixin26@kstu.kg

I. Razzakov Kyrgyz State Technical University

Bojian Guo

I. Razzakov Kyrgyz State Technical University

Asel Kartanova

I. Razzakov Kyrgyz State Technical University

Research Article

Keywords: agricultural decision support, dataset shift, human-in-the-loop, small-data forecasting, temporal validation, wheat yield

Posted Date: September 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10698325/v1

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

Additional Declarations: No competing interests reported.

Abstract

Purpose

Small annual agricultural panels are vulnerable to validation leakage and silent performance loss when models are transferred to later years or unobserved regions. This study evaluates an auditable decision-support workflow that freezes model routing and review capacity before external testing.

Methods

Wheat yields for seven Kyrgyz oblasts (1991–2025) were combined with lagged temperature context. Development covered 1994–2014, calibration 2015–2018, and sealed evaluation 2019–2025 in five seen and two held-out oblasts. Persistence, three learned models, and calibration-frozen adaptive fallback routing (CAFR) were assessed using exact year-cluster inference, 21 spatial holdouts, interval stress tests, and enumeration of 7^7 equal-capacity review policies.

Results

Row-random validation understated future CAFR error by 30.2%. Across 49 predictions, random forest had the lowest pooled mean absolute error (3.43 centner ha − 1), followed by CAFR (3.54), ridge (3.56), and persistence (4.16). CAFR improved on persistence in four years, tied in three, and worsened in none; the one-sided exact p value was 0.0625. A nominal 90% frozen interval covered only 55.1% of external cases. Reviewing one highest-drift oblast per year examined 14.3% of cases and captured 33.0% of absolute error. Only 3 of 823,543 equal-capacity random policies matched or exceeded that capture ($p = 3.64 \times 10 - 6$).

Conclusion

The contribution is a reproducible reliability audit and workload-control protocol, not evidence for autonomous yield forecasting. Exact small-sample inference, failure reporting, and bounded human review provide a defensible basis for sparse-data agricultural decision support.

1 Introduction

Digital agriculture increasingly relies on information systems that transform production records, environmental observations, and management data into forecasts, alarms, and priorities for inspection. In biosystems engineering, the useful product is therefore more than a point estimate. A deployable system needs a traceable data path, a stated operating range, and a predefined response when new observations differ from the conditions used for development. Recent Journal of Biosystems Engineering studies illustrate this systems perspective through open-source crop-monitoring software

and machine-learning workflows connected to measurable agricultural or biological tasks (Kim, 2024; Aliee et al., 2026). The present study examines a related reliability problem: what can be defended when annual regional data are sparse and operational conditions change over time and space?

The Kyrgyz Republic provides a stringent small-data setting. Wheat production is reported for seven oblasts with contrasting topography and production conditions. National and regional assessments identify warming, drought, water stress, and uneven agricultural exposure as material constraints (World Bank Group & Asian Development Bank, 2021; Intergovernmental Panel on Climate Change, 2022; Park et al., 2021). Long-term adaptation analysis is essential, but administrative decisions also recur each year: whether a baseline estimate is credible, whether a regional record requires verification, and where limited agronomic investigation should be concentrated.

Crop-yield modelling has advanced through process-based simulation, weather and management covariates, remote sensing, unmanned aerial systems, and machine learning (Lobell et al., 2011; Challinor et al., 2014; Crane-Droesch, 2018; van Klompenburg et al., 2020; Shahhosseini et al., 2021). Recent regional wheat studies using Sentinel-2 or comprehensive UAV data show what becomes possible when crop-season and spatially resolved information is available (Khodjaev et al., 2025; Meng et al., 2026). By contrast, administrative panels often contain only annual outcomes and coarse context. In that setting, sophisticated models can still produce apparently precise results, but the evidence boundary must be narrower.

Model rankings are highly sensitive to the validation boundary. Random partitions may place adjacent years and related regions on both sides of a split, making interpolation look like forecasting. Blocked temporal and spatial validation is consequently recommended for structured environmental and agricultural data (Tashman, 2000; Roberts et al., 2017; Bergmeir et al., 2018). Spatial validation has exposed weak transfer in ecological mapping (Ploton et al., 2020), and recent crop-yield work has shown that data partitioning can materially change apparent accuracy (Petrovic et al., 2026). Leakage is not merely a statistical nuisance; it can make a computational system appear ready for use when it has not been tested in the intended operating regime (Kapoor & Narayanan, 2023).

A second engineering problem is silent degradation after dataset shift (Moreno-Torres et al., 2012; Gama et al., 2014; Rabanser et al., 2019). Selective prediction and human-in-the-loop approaches respond by escalating cases that do not satisfy a reliability criterion instead of treating all outputs as equally safe (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Mosqueira-Rey et al., 2023). Conformal prediction can convert calibration residuals into interpretable intervals, but its finite-sample guarantee depends on exchangeability and can weaken under temporal change (Vovk et al., 2005; Lei et al., 2018; Gibbs & Candes, 2021; Angelopoulos & Bates, 2023). A practical agricultural information system should therefore report not only error, but also interval failure, routing harm, and the workload created by review. More broadly, predictive uncertainty can degrade under dataset shift, so uncertainty estimates require explicit evaluation outside the development distribution (Ovadia et al., 2019).

The research question was whether a development- and calibration-frozen fallback rule, coupled to a fixed annual review capacity, could provide a more auditable one-year-ahead wheat-yield baseline under temporal and spatial shift. The study contributes: (1) a direct audit of row-random validation optimism; (2) a frozen routing layer between persistence and ridge regression; (3) exact year-cluster inference that respects the seven external years; (4) explicit accounting of beneficial and harmful routes and interval failures; and (5) exact evaluation of a one-case-per-year review policy against all equal-capacity random allocations. The complete workflow was repeated for all 21 possible two-oblast holdouts. Negative results were retained because they define the system boundary.

2 Materials and Methods

2.1 Decision context and study unit

The target was annual wheat yield in each oblast in target year t, reported in centner per hectare (1 centner ha − 1 = 100 kg ha − 1). The intended use is inter-annual administrative screening after official information for t − 1 becomes available. It is not an in-season crop forecast and does not prescribe field-level action. No target-year yield, target-year climate value, future revision, or test-derived quantity was allowed as a predictor.

The primary spatial holdout comprised Batken and Naryn, representing contrasting south-western and high-elevation interior settings. Chuy, Issyk-Kul, Jalal-Abad, Osh, and Talas were treated as seen regions. Because a single geographic split can be fortuitous, the frozen procedure was later repeated for every pair of held-out oblasts. Those 21 configurations are sensitivity settings built from the same seven regions and seven target years; they are not 21 independent external studies.

2.2 Data sources and leakage-safe features

The first source was version 1.1 of the National Statistical Committee of the Kyrgyz Republic data set Crops yield by territory (National Statistical Committee of the Kyrgyz Republic, 2026). It contained complete annual wheat-yield values for seven oblasts from 1991 through 2025, giving 245 raw region-year observations. The downloaded file was retained unchanged and its SHA-256 hash was recorded.

The second source was the national annual land-surface temperature anomaly distributed by Our World in Data from modified Copernicus ERA5 information (Hersbach et al., 2020; Ritchie et al., 2024). It represents the annual difference from the 1991–2020 national mean. A national annual anomaly cannot represent local crop-season exposure or separate heat, precipitation, snowpack, irrigation, and phenological effects. It was used only as lagged background context and was not interpreted causally. Stage-specific wheat research demonstrates why seasonal extremes and crop development matter, and why the present variable cannot replace local agronomic measurements (Schierhorn et al., 2021).

For each target year, eight predictors were calculated only from information available through t − 1: previous yield, three-year mean yield, three-year yield standard deviation, three-year linear yield slope,

previous national temperature anomaly, three-year anomaly mean, three-year anomaly standard deviation, and a centred year index. Three lag years were required, so the modelling panel began in 1994 and contained 224 observations. No missing-value imputation, interpolation, outlier deletion, or winsorization was applied.

2.3 Frozen development, calibration, and external evaluation

The development set covered 1994–2014 in five seen oblasts (n = 105). Calibration covered 2015–2018 in those oblasts (n = 20). The later-time test covered 2019–2025 in the same five oblasts (n = 35), and the simultaneous spatial test covered the same years in Batken and Naryn (n = 14). Thus, 49 external prediction rows were attached to only seven independent target years. Standardization, robust feature centres, interquartile ranges, and model fitting used development data only. The routing threshold and residual interval radii used calibration data only. No setting was changed after external outcomes were calculated. The frozen partitions and their decision roles are summarized in Table 1.

Table 1
Frozen data partitions and decision roles

Component	Years	Observations	Role	Information control
Raw source	1991–2025	7 oblasts × 35 years = 245	Source preservation and descriptive context	Modelling starts in 1994 because three lag years are required
Development	1994–2014	5 seen oblasts; n = 105	Fit models and development feature statistics	All transformations frozen here
Calibration	2015–2018	5 seen oblasts; n = 20	Select routing threshold and residual interval radii	No external outcomes used
Later-time test	2019–2025	5 seen oblasts; n = 35	Transfer to later years	Sealed evaluation
Spatial test	2019–2025	Batken and Naryn; n = 14	Same-time transfer to unseen regions	Sealed evaluation

Note. All target-year predictors use information available through t − 1.

2.4 Predictors, comparators, and calibration-frozen routing

Persistence predicted the current yield from the preceding observed yield. Ridge regression used standardized features and a fixed penalty of 10 (Hoerl & Kennard, 1970). Random forest used 120 trees, maximum depth five, minimum leaf size three, and 80% feature sampling (Breiman, 2001). Gradient boosting used 100 depth-two trees, learning rate 0.04, minimum leaf size three, and Huber loss

(Friedman, 2001). These deliberately conservative specifications were fixed rather than chosen through a large search that could adapt to a small test set.

Calibration-frozen adaptive fallback routing (CAFR) was a decision layer, not a new regression estimator. A robust drift score measured multivariate distance from the development operating envelope. For case i with p lagged features, development median m_j, and interquartile range q_j, the score was:

$$d_i \;=\; sqrt\,[(\,1/p)\; sum_(j=1)^{p}\; ((x_ij \;-\; m_j)/q_j)^2]$$

1

Near-zero interquartile ranges were replaced by one. The score is an operational distance, not a probability, and it does not attribute shift to temperature alone. Candidate routing thresholds were the 5th through 95th percentiles of calibration scores in five-percentage-point steps. The selected threshold minimized calibration MAE plus 0.2 times the calibration 90th-percentile absolute error. Persistence was retained at or below the threshold; higher-drift cases were routed to ridge regression:

$$yhat_i\,(CAFR) \;=\; yhat_i\,(ridge)\;\; if\; d_i \;>\; tau;\; y_i, t-1\; otherwise$$

2

The selected threshold, τ = 1.228, was fixed before the future and spatial tests.

2.5 Capacity-constrained review and interval stress test

A second layer ranked the seven oblast predictions within each target year by the same frozen drift score. The primary policy sent exactly one highest-drift oblast to human review, giving a fixed review rate of one seventh and automatic coverage of six sevenths. The flag did not use the target outcome. Review was defined as a request for supplementary local evidence—such as crop-season heat and precipitation, irrigation delivery, planted area, field reports, pest observations, or statistical revisions—before an administrative decision. Capacities of zero, two, and three cases per year were retained as sensitivity analyses and were not selected from external performance.

Calibration absolute residuals were also used to construct split-conformal intervals at nominal 80% and 90% levels. The finite-sample residual rank and radius were fixed on the 20 calibration cases and applied unchanged to external predictions. Because the external observations occurred later and partly in unseen regions, this was an interval-transport stress test rather than an assertion that exchangeability held. We report empirical coverage, coverage gap, misses, and the proportion of misses captured by the fixed review rule.

2.6 Evaluation, exact inference, and sensitivity analysis

Performance measures were MAE, root mean squared error, signed mean error, the 90th percentile of absolute error, and the largest oblast-specific MAE. Future and spatial tests were reported separately

and pooled. Random-validation optimism was assessed with ten fixed stratified row-random seeds that matched per-region sample counts while intentionally ignoring chronology.

The seven target years, rather than the 49 region-year rows, were the independent external units. For each comparator, the annual mean absolute-error difference was calculated and all 2^7 sign reversals were enumerated. One-sided and two-sided exact p values were obtained from that randomization distribution. A paired year-cluster bootstrap with 5,000 fixed-seed replicates was retained only as a descriptive interval, not as the primary significance claim (Efron & Tibshirani, 1993).

The fixed review rule was evaluated by exact enumeration of all 7^7 = 823,543 policies that select one of seven oblasts independently in each of seven years. Its error-capture share was compared with this complete equal-capacity randomization distribution. Sensitivity analyses repeated the full frozen procedure for all 21 two-oblast holdouts, four history lengths, threshold multipliers from 0.8 to 1.2, and leave-one-calibration-year-out threshold estimation.

2.7 Reproducibility and language-model assistance

Online Resource 1 contains the public raw files, processed panel, frozen split files, source hashes, data dictionary, fixed configuration, deterministic Python code, automated tests, case-level predictions, interval and review audits, exact-test outputs, and figure data. The package follows reproducible and FAIR research principles (Peng, 2011; Wilkinson et al., 2016). A clean CPU run regenerates the reported outputs and passes 13 automated tests.

An OpenAI language model assisted with language revision, translation, code inspection, document assembly, and consistency checking. It did not create observations or independently generate numerical findings. Every numerical statement was regenerated by the included deterministic workflow and checked against machine-readable outputs. The authors verified the sources, design, interpretation, and final text and retain full responsibility.

3 Results

3.1 Shift and validation-design optimism

Regional wheat yields were heterogeneous over 1991–2025, with persistent periods and abrupt reversals (Fig. 1). The national annual temperature anomaly became predominantly positive later in the record and reached 2.07°C in 2025. This shows that the external period did not simply reproduce the early background; it does not establish a temperature–yield causal effect.

Across ten row-random seeds, median CAFR MAE was 2.89 centner ha − 1, compared with 4.15 in the sealed future test. Row-random validation therefore understated later CAFR error by 30.2%. Understatement ranged from 35.1% to 40.5% for persistence and the three always-on learned models (Fig. 2).

3.2 Later-time and unseen-region performance

Persistence was strongest in calibration, with MAE 1.64 centner ha − 1; CAFR routed 15% of calibration cases and achieved 1.22. The ordering changed in later years. In the five seen oblasts, persistence, ridge, random forest, gradient boosting, and CAFR achieved MAEs of 4.91, 4.27, 3.99, 3.87, and 4.15, respectively. CAFR was not the best later-time model, but it reduced the large persistence failure and kept signed bias close to zero.

In Batken and Naryn, ridge achieved MAE 1.79, CAFR 2.01, random forest 2.04, persistence 2.30, and gradient boosting 3.36. Across all 49 external predictions, random forest had the lowest average error at 3.43, followed by CAFR at 3.54 and ridge at 3.56 (Table 2). CAFR reduced persistence error by 15.1%, reduced the largest oblast-specific MAE from 11.24 to 7.56, and changed pooled signed bias from 0.27 to − 0.04 centner ha − 1.

Table 2
External predictive performance in centner per hectare

Method	Later-time MAE	Spatial MAE	Pooled MAE	Pooled bias	Largest-oblast MAE
Persistence	4.91	2.30	4.16	0.27	11.24
Ridge regression	4.27	1.79	3.56	−1.27	7.94
Random forest	3.99	2.04	3.43	−0.73	7.67
Gradient boosting	3.87	3.36	3.72	0.76	7.71
CAFR	4.15	2.01	3.54	−0.04	7.56

3.3 Year-cluster evidence and routing safety

CAFR improved on persistence in four target years, tied it in three, and was worse in none. The observed difference in annual-cluster-weighted MAE was − 0.63 centner ha − 1. Exact enumeration of all 2^7 sign reversals gave a one-sided p value of 0.0625 and a two-sided p value of 0.125. The year-cluster bootstrap interval was − 1.16 to − 0.17, but the exact test is more transparent for only seven years. The data support a consistent direction in this evaluation window, not a general proof of superiority.

CAFR routed 19 of 49 external cases to ridge. Eleven switches reduced absolute error and eight increased it, giving a beneficial-switch rate of 57.9%. Helpful switches improved error by 1.62 centner ha − 1 on average; harmful switches cost 1.47. The routing layer improved aggregate reliability while remaining unsafe for a substantial minority of switched cases.

3.4 Interval failure and exact review-allocation audit

Calibration-frozen residual intervals did not retain nominal coverage. For CAFR, nominal 80% and 90% intervals achieved only 51.0% and 55.1% pooled external coverage. The nominal 90% radius was 3.09

centner ha − 1, and 22 of 49 observations fell outside it. Historical calibration therefore did not repair temporal and spatial shift.

The primary review policy examined seven cases (14.3%). The 42 automatic cases had CAFR MAE 2.76, compared with 8.18 for reviewed cases. Review captured 33.0% of total absolute error, a lift of 2.31 over the 14.3% expected under uniform allocation. Among all 823,543 policies selecting one random oblast per year, only three matched or exceeded the observed capture (exact $p = 3.64 \times 10 - 6$). The central 95% range under random allocation was 6.5%–23.8% error capture (Fig. 3).

All seven reviewed cases were nominal 90% interval misses, but automatic-case interval coverage remained only 64.3%. Review prioritization therefore concentrated attention on difficult cases; it did not validate the frozen interval. Chuy was selected in five of seven years, so the policy also showed concentration risk that should be monitored prospectively.

3.5 Spatial and calibration sensitivity

CAFR improved on persistence in all 21 possible two-oblast holdout configurations. Relative improvement ranged from 6.3% to 20.8%, with a median of 13.9%; CAFR was nevertheless the absolute best model in only two configurations. These overlapping sensitivity settings support stable baseline repair, not 21 independent replications or universal superiority (Fig. 4).

The one-case review lift ranged from 2.02 to 2.56 across the 21 configurations, and its error-capture share ranged from 28.9% to 36.6%. The most frequently selected oblast accounted for 57.1%–71.4% of review flags, confirming that operational monitoring should include geographic concentration.

Leaving one calibration year out produced thresholds from 1.194 to 1.321 and pooled external CAFR MAEs from 3.42 to 3.58. The main threshold was therefore not determined by a single calibration year. A multiplier of 0.8 would have lowered pooled error after inspection, but it was not chosen by the frozen protocol and remains a post hoc result. The principal reliability, uncertainty, and failure audits are summarized in Table 3.

Table 3
Reliability, uncertainty, and failure audits

Audit	Quantity	Result	Interpretation
Row-random validation	Future CAFR error understated	30.2%	Ten fixed stratified row-random seeds
Year-cluster direction	Years CAFR better / equal / worse than persistence	4 / 3 / 0	Exact one-sided p = 0.0625; two-sided p = 0.125
Routing harm	Harmful ridge routes among all routes	8/19 (42.1%)	Negative result retained
Nominal 90% interval	Pooled external coverage	55.1%	Frozen calibration radius = 3.09 centner ha − 1
Fixed review capacity	Review rate; total error captured	14.3%; 33.0%	One highest-drift oblast per year
Exact review randomization	Equal-capacity policies matching/exceeding observed capture	3/823,543	Exact p = 3.64 × 10 − 6
Spatial sensitivity	Holdout settings where CAFR beat persistence	21/21	Overlapping settings, not independent replications
Review sensitivity	Capture lift across 21 holdout settings	2.02–2.56	Most-selected-region share = 57.1%–71.4%
Calibration stability	Leave-one-year threshold; external MAE	1.194–1.321; 3.42–3.58	No single calibration year determined the main result

4 Discussion

4.1 Validation design as an engineering control

Model rankings reversed between calibration, later years, and unseen regions. The ten-seed audit quantified the consequence: random splitting understated future CAFR error by 30.2%, and the understatement approached 40% for the tree ensembles. This finding is consistent with work showing that random folds can exaggerate transfer in structured ecological and agricultural data (Roberts et al., 2017; Ploton et al., 2020; Petrovic et al., 2026). In an agricultural information system, validation is not a reporting afterthought. It is a design control that determines which operating regime an accuracy claim describes.

The exact sign-flip audit also changes the interpretation. Treating 49 rows as independent would create artificial precision; treating seven years as independent yields a one-sided p value of 0.0625. The result remains practically relevant—four annual improvements and no annual deterioration—but is statistically

fragile. This is precisely the kind of distinction that should be visible before a sparse-data system is deployed.

4.2 Relation to biosystems engineering and yield-prediction literature

JBE covers information technology, precision agriculture, instrumentation and control, smart farming, artificial intelligence, and big-data analytics as engineering components of living systems. Within that frame, open-source satellite crop monitoring demonstrates the value of transparent data processing and usable software outputs (Kim, 2024), while controlled sensing studies connect machine learning to measurable biological tasks (Aliee et al., 2026). The present work contributes a reliability layer for a different data regime: sparse official records, annual decisions, and constrained capacity for follow-up.

The modest predictive ceiling is not surprising. Regional deep-learning studies using Sentinel-2 and environmental variables, and Central Asian UAV studies with richer crop observations, operate with information that is unavailable in the present panel (Khodjaev et al., 2025; Meng et al., 2026). Weather effects on wheat also depend on developmental stage and extremes, not only annual national anomalies (Schierhorn et al., 2021). The current system should therefore be viewed as an auditable administrative baseline and triage mechanism, not a substitute for agronomic sensing.

4.3 Simple baselines, learned models, and routing trade-offs

Random forest had the lowest pooled row-average MAE, and gradient boosting led in the seen-region future test. Those results prevent the fallback rule from being described as the best predictor. CAFR instead offered near-zero pooled bias, a smaller worst-oblast error than persistence, transparent route provenance, and a simple default action when drift was low. Whether those properties justify deployment depends on the cost assigned to bias, extreme regional error, maintenance, and human review.

The 42.1% harmful-switch rate is a particularly important negative result. A distance score can identify unfamiliar inputs without reliably identifying which model will be better. The routing score should therefore be treated as an escalation signal rather than as proof that ridge is superior. Future work should compare cost-sensitive gates, online recalibration, and model-specific uncertainty while preserving a prospective evaluation boundary.

4.4 Interval failure and human review

Split-conformal prediction is attractive because it converts residuals into distribution-free intervals under exchangeability (Vovk et al., 2005; Lei et al., 2018; Angelopoulos & Bates, 2023). Later years and unseen regions need not satisfy that condition. Adaptive conformal methods address distribution shift in larger streams (Gibbs & Candes, 2021), but four calibration years provide little capacity for stable adaptation. The collapse of nominal 90% coverage to 55.1% is therefore not a peripheral defect; it is evidence that a familiar uncertainty wrapper cannot be transported uncritically.

The review audit offers a different kind of control. Human-in-the-loop systems are most useful when the human role, trigger, and workload are explicit (Mosqueira-Rey et al., 2023). Here, one predefined case per year captured substantially more error than any but three of 823,543 equal-capacity random policies. Yet no expert actually corrected a prediction, and review concentration in Chuy was high. The result validates prioritization in retrospect, not the effectiveness, cost, or fairness of a real review process.

4.5 Agricultural interpretation, limitations, and future work

Kyrgyzstan's agricultural exposure is spatially differentiated. Cropland-suitability research and national climate assessments point to regional variation in terrain, warming, drought, and water-system pressure (Park et al., 2021; World Bank Group & Asian Development Bank, 2021). The present annual model is far coarser than field- or season-scale systems. Its defensible use is screening: when a case is flagged, planners should inspect local precipitation and heat, irrigation delivery, planted area, cultivars, pests, field observations, and statistical revisions. The model cannot diagnose local crop stress or support farm-level prescriptions. This caution is consistent with global evidence that climate variability explains a substantial but incomplete share of crop-yield variability, leaving important non-climatic determinants outside climate-only models (Ray et al., 2015).

Seven oblasts and seven external years provide few independent temporal units. Annual yield combines weather, technology, seed, irrigation, area, markets, and statistical practice. The spatial test remains inside one national statistical system, and the review rule has not been prospectively evaluated by an agency. No editing or algorithmic complexity can replace these missing observations.

A deployed prediction service is a maintained information system. Changes in data definitions, software dependencies, thresholds, and undocumented manual decisions create technical debt (Sculley et al., 2015). The supplied resource records raw hashes, data versions, split files, feature statistics, case-level predictions, interval failures, review flags, exact tests, figure data, requirements, and a one-command run. The next study should add crop-season oblast-level weather, snow and soil moisture, irrigation, planted area, phenology, and remote sensing; preregister recalibration and stopping rules; assign explicit costs to false reassurance and unnecessary review; evaluate whether reviewers can correct flagged cases; and test the workflow in another Central Asian statistical system.

5 Conclusions

A sparse annual wheat-yield decision-support workflow was evaluated under later time and simultaneous spatial transfer rather than random historical interpolation. Row-random validation understated future CAFR error by 30.2%. Across 49 sealed external predictions, random forest had the lowest pooled average MAE, while CAFR reduced persistence MAE from 4.16 to 3.54 centner ha − 1, reduced the largest oblast error, and nearly removed signed bias. CAFR improved in four external years and tied in three; the exact one-sided p value was 0.0625, so the evidence should not be overstated.

Uncertainty did not transfer automatically. A nominal 90% frozen interval covered only 55.1% of external observations. Reviewing one of seven oblasts per year concentrated 33.0% of total absolute error into 14.3% of cases, and this capture was exceeded by only three of 823,543 equal-capacity random policies. The defensible contribution is an auditable reliability and workload-control protocol that exposes validation optimism, interval failure, route harm, and review concentration before sparse administrative predictions are trusted.

Declarations

Data availability: The analysis-ready datasets supporting this study, together with data provenance, variable definitions, licensing information, data-audit documentation, frozen analysis configuration, reproducibility tests, and key analysis outputs, are publicly available in the GitHub repository: https://github.com/hezejuyede/wheat-yield-reliability-audit. The complete reproducibility package is also provided as Online Resource 1 and contains the public raw files, processed panel, frozen splits, source hashes, deterministic Python code, fixed configuration, automated tests, predictions, exact inference, interval and review audits, and figure data.

Code availability: The complete CPU workflow and pinned software requirements are included in Online Resource 1. Running python 04_CODE/run_all.py regenerates the outputs and executes 13 automated tests.

Ethics approval: Not applicable. The study used aggregated public agricultural and meteorological statistics and involved no human participants, animals, or personally identifiable information.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Clinical trial number: not applicable.

Author Contribution

X.L. conceived and designed the study, developed the methodology, curated and analysed the data, implemented and validated the computational workflow, prepared all figures and tables, and wrote the original manuscript. B.G. contributed to validation, formal analysis, and critical revision of the manuscript. A.K. supervised the research, reviewed the methodology, and critically revised the manuscript. All authors reviewed and approved the final version of the manuscript.

References

1. Aliee M, Marateb HR, Vasefi F, et al. Fluorescence imaging and machine learning for surface contamination detection: A comparative study with ATP testing. J Biosystems Eng. 2026;51:9.

https://doi.org/10.1007/s42853-025-00293-w.

2. Angelopoulos AN, Bates S. Conformal prediction: A gentle introduction. Found Trends Mach Learn. 2023;16(4):494–591. https://doi.org/10.1561/2200000101.
3. Bergmeir C, Hyndman RJ, Koo B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. Comput Stat Data Anal. 2018;120:70–83. https://doi.org/10.1016/j.csda.2017.11.003.
4. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
5. Challinor AJ, Watson J, Lobell DB, Howden SM, Smith DR, Chhetri N. A meta-analysis of crop yield under climate change and adaptation. Nat Clim Change. 2014;4:287–91. https://doi.org/10.1038/nclimate2153.
6. Crane-Droesch A. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environ Res Lett. 2018;13(11):114003. https://doi.org/10.1088/1748-9326/aae159.
7. Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman & Hall; 1993.
8. El-Yaniv R, Wiener Y. On the foundations of noise-free selective classification. J Mach Learn Res. 2010;11:1605–41.
9. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals Stat. 2001;29(5):1189–232. https://doi.org/10.1214/aos/1013203451.
10. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM-CSUR. 2014;46(4):44. https://doi.org/10.1145/2523813.
11. Geifman Y, El-Yaniv R. (2017). Selective classification for deep neural networks. Adv Neural Inf Process Syst, 30.
12. Gibbs I, Candès E. Adaptive conformal inference under distribution shift. Adv Neural Inf Process Syst. 2021;34:1660–72.
13. Hersbach H, Bell B, Berrisford P, et al. The ERA5 global reanalysis. Q J R Meteorol Soc. 2020;146(730):1999–2049. https://doi.org/10.1002/qj.3803.
14. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634.
15. Intergovernmental Panel on Climate Change. Climate change 2022: Impacts, adaptation and vulnerability. Cambridge University Press; 2022. https://doi.org/10.1017/9781009325844.
16. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. https://doi.org/10.1016/j.patter.2023.100804.
17. Khodjaev S, Bobojonov I, Kuhn L, Glauben T. Model Earth Syst Environ. 2025;11:15. https://doi.org/10.1007/s40808-024-02188-9. Optimizing machine learning models for wheat yield estimation using a comprehensive UAV dataset.
18. Kim JY. Open-source software for satellite-based crop health monitoring. J Biosystems Eng. 2024;49:419–33. https://doi.org/10.1007/s42853-024-00242-z.

19. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. J Am Stat Assoc. 2018;113(523):1094–111. https://doi.org/10.1080/01621459.2017.1307116.
20. Lobell DB, Schlenker W, Costa-Roberts J. Climate trends and global crop production since 1980. Science. 2011;333(6042):616–20. https://doi.org/10.1126/science.1204531.
21. Meng H, Segarra J, Wang Z, Araus JL, Kefauver SC. Assessing deep learning models for regional wheat yield prediction using Sentinel-2 and environmental data. Precision Agric. 2026;27:109. https://doi.org/10.1007/s11119-026-10415-9.
22. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern Recogn. 2012;45(1):521–30. https://doi.org/10.1016/j.patcog.2011.06.019.
23. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: A state of the art. Artif Intell Rev. 2023;56:3005–54. https://doi.org/10.1007/s10462-022-10246-w.
24. National Statistical Committee of the Kyrgyz Republic. (2026). Crops yield by territory (Version 1.1) [Data set]. https://stat.gov.kg/ru/statistics/download/dynamic/1272/
25. Ovadia Y, Fertig E, Ren J et al. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Adv Neural Inf Process Syst, 32.
26. Park S, Lim CH, Kim SJ, Isaev E, Choi SE, Lee SD, et al. Assessing climate change impact on cropland suitability in Kyrgyzstan: Where are potential high-quality cropland and the way to the future. Agronomy. 2021;11(8):1490. https://doi.org/10.3390/agronomy11081490.
27. Peng RD. Reproducible research in computational science. Science. 2011;334(6060):1226–7. https://doi.org/10.1126/science.1213847.
28. Petrovic B, Paluba D, Nofrizal AY, Kučera A. Analyzing meteorological effects on crop yield and yield prediction through machine learning with different data partitioning approaches: A case study from Czech Republic. J Agric Food Res. 2026;26:102662. https://doi.org/10.1016/j.jafr.2026.102662.
29. Ploton P, Mortier F, Réjou-Méchain M, et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat Commun. 2020;11:4540. https://doi.org/10.1038/s41467-020-18321-y.
30. Rabanser S, Günnemann S, Lipton ZC. Failing loudly: An empirical study of methods for detecting dataset shift. Adv Neural Inf Process Syst. 2019;32:1394–406.
31. Ray DK, Gerber JS, MacDonald GK, West PC. Climate variation explains a third of global crop yield variability. Nat Commun. 2015;6:5989. https://doi.org/10.1038/ncomms6989.
32. Ritchie H, Rosado P, Samborska V. (2024). Annual temperature anomalies [Data set]. Our World in Data. https://ourworldindata.org/grapher/annual-temperature-anomalies
33. Roberts DR, Bahn V, Ciuti S, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913–29. https://doi.org/10.1111/ecog.02881.

34. Schierhorn F, Hofmann M, Gagalyuk T, Ostapchuk I, Müller D. Machine learning reveals complex effects of climatic means and weather extremes on wheat yields during different plant developmental stages. Clim Change. 2021;169:39. https://doi.org/10.1007/s10584-021-03272-0.
35. Sculley D, Holt G, Golovin D et al. (2015). Hidden technical debt in machine learning systems. Adv Neural Inf Process Syst, 28.
36. Shahhosseini M, Hu G, Huber I, Archontoulis SV. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Sci Rep. 2021;11:1606. https://doi.org/10.1038/s41598-020-80820-1.
37. Tashman LJ. Out-of-sample tests of forecasting accuracy: An analysis and review. Int J Forecast. 2000;16(4):437–50. https://doi.org/10.1016/S0169-2070(00)00065-0.
38. van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: A systematic literature review. Comput Electron Agric. 2020;177:105709. https://doi.org/10.1016/j.compag.2020.105709.
39. Vovk V, Gammerman A, Shafer G. (2005). Algorithmic learning in a random world. Springer. https://doi.org/10.1007/b106715
40. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18. The FAIR guiding principles for scientific data management and stewardship.
41. World Bank Group, & Asian Development Bank. (2021). Climate risk country profile: Kyrgyz Republic. World Bank. https://doi.org/10.1596/36377

Figures

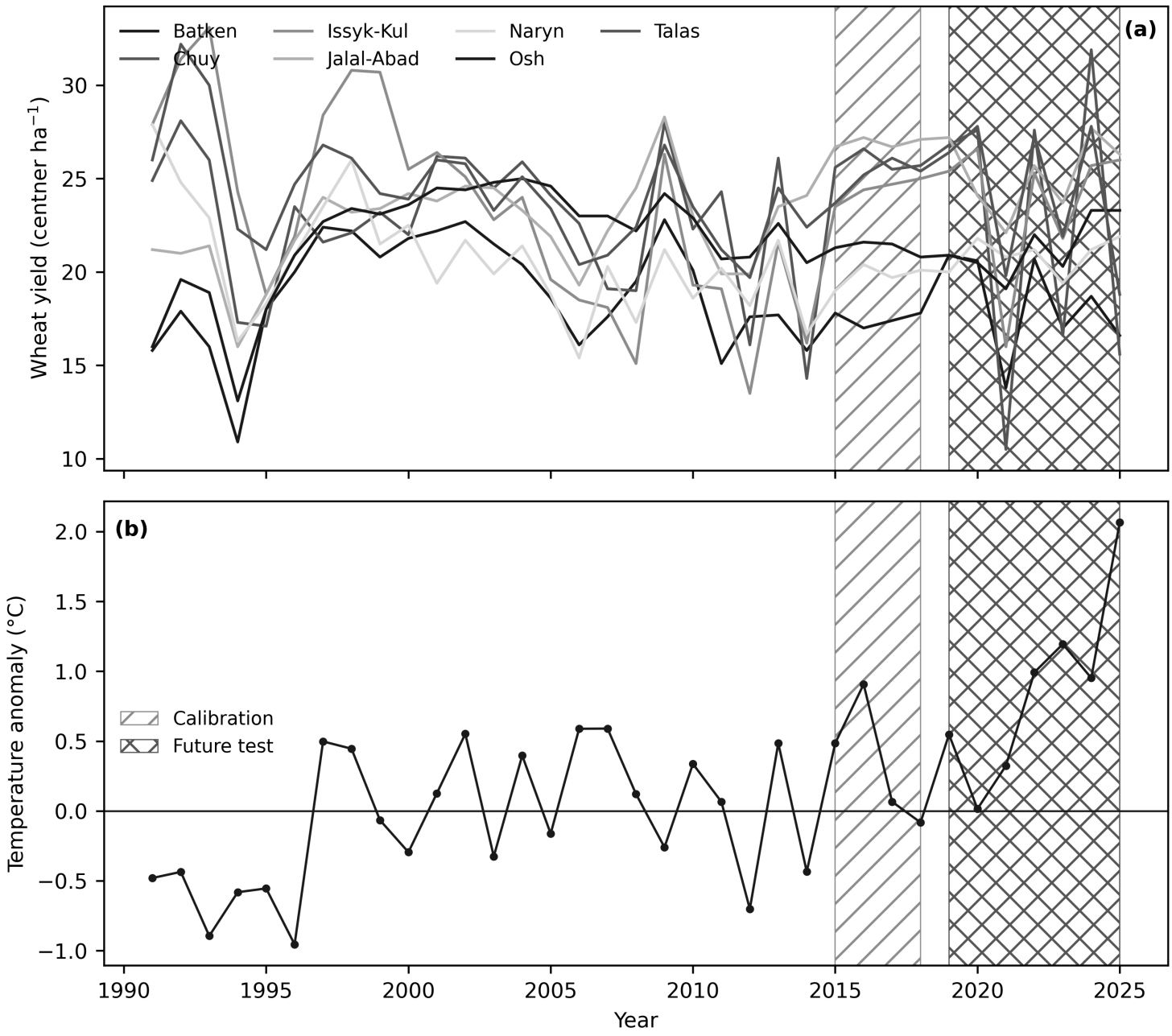


Figure 1

Regional wheat-yield histories, lagged national annual temperature anomaly, and frozen evaluation windows

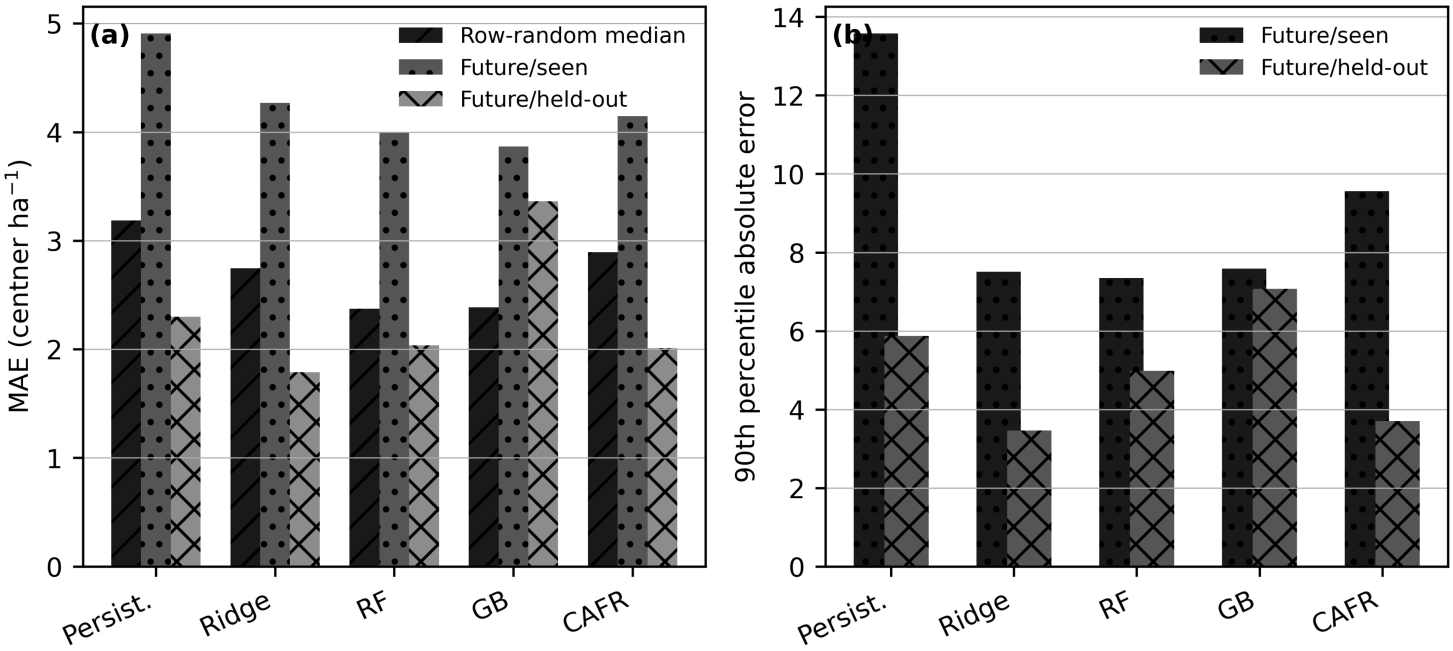


Figure 2

Validation-design audit and external predictive performance

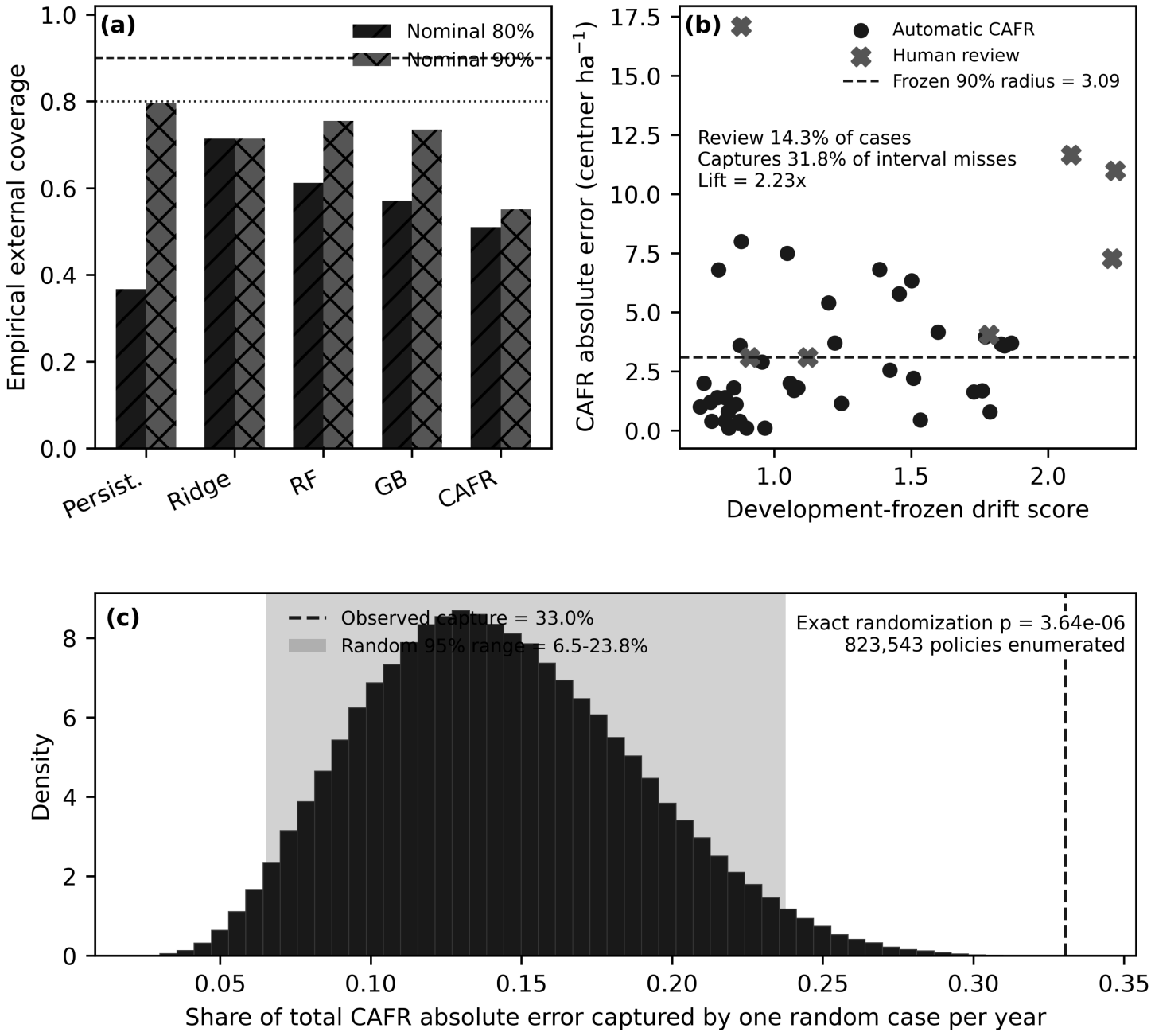


Figure 3

Calibration-frozen interval stress, drift-based review, and exact equal-capacity randomization distribution

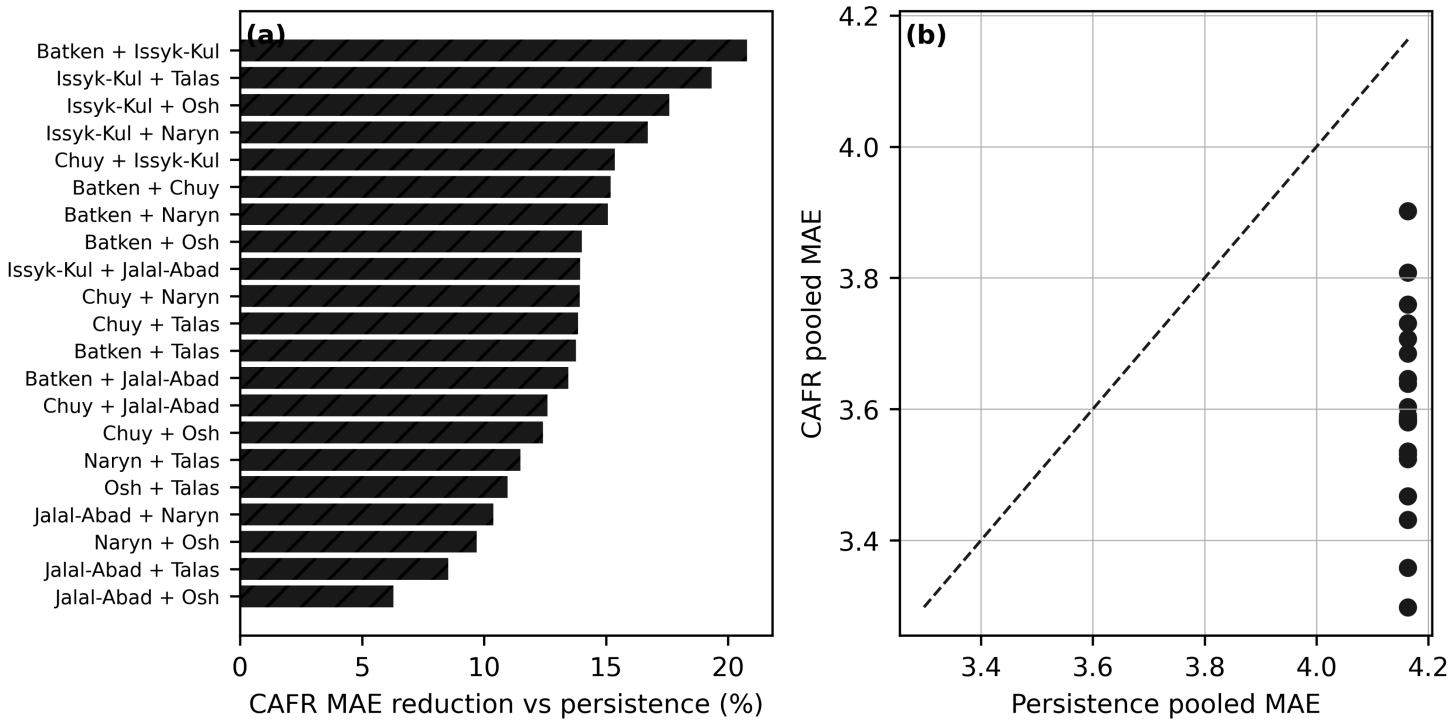


Figure 4

External performance and review sensitivity across all 21 two-oblast spatial holdouts

Supplementary Files

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

- 02OnlineResource1AnonymousReproducibilityPackage.zip