

2 **Supplementary information**

3 Phenotyping part of the target family improves within-family genomic prediction in rice nested
4 association mapping families

5 Jixiao Li

6 Field Science Center for Northern Biosphere, Hokkaido University, Sapporo 060-0811, Japan

7 Correspondence: yuriljx@gmail.com

8 This supplement contains detailed designs, analysis chronology, complete result-file definitions and
9 all supplementary figures. Canonical result tables retain full precision. U and ΔU are g/plant for
10 NAM/MAGIC and kg/ha for IRRI; predictive correlations are dimensionless. Narrative values
11 generally use three decimals in g/plant, one decimal in kg/ha and three decimals for correlations,
12 with extra precision when rounding would conceal a nonzero sign. References cited in this file are
13 listed in the main article.

14 **S1. Source records, genotype coding and phenotype boundaries**

15 Nested association mapping (NAM) and multiparent advanced generation intercross (MAGIC)
16 materials originate from Wei et al. (2024); family identities follow Chen (2026;
17 <https://doi.org/10.6084/m9.figshare.31881931.v1>). The NAM resource consists of recombinant
18 inbred families from the elite indica common parent Huanghuazhan crossed with 15 other diverse
19 accessions, followed by seven rounds of single-seed descent (Wei et al. 2024). The source
20 experiment used single-row plots of seven plants in an augmented block design, with each family
21 blocked separately and Huanghuazhan included as a check (Wei et al. 2024). Data S4 supplies the
22 analyzed grain-yield columns K and Z under Shanghai and Hangzhou, headed Grain yield (g). The
23 primary paper supports grain yield per plant, but the specific K/Z field year and the transformation
24 from plot measurements to released values, including aggregation, block adjustment or best linear
25 unbiased estimate (BLUE) calculation, remain unresolved. Historical year labels are not an
26 independently established K/Z mapping. Data S12 records three grain-yield samples per plot for

27 18K-rice at SH/HZ, not three independent field replicates or evidence that Data S4 is their arithmetic
28 mean; parental sampling counts differ.

29 The NAM historical panel contains 2048 markers selected using all 16019 NAM genotypes without
30 phenotype data. Filters were call rate ≥ 0.95 , heterozygous fraction among called genotypes ≤ 0.05
31 and allele-copy minor-allele frequency (MAF) ≥ 0.05 . Cached genotype codes 0/1/2/3 map to dosages
32 0/2/1/missing. If n_0 , n_1 and n_2 count the called cached codes, allele frequency and MAF are:

33
$$p = \frac{2n_1 + n_2}{2(n_0 + n_1 + n_2)}$$

34
$$MAF = \min(p, 1 - p)$$

35 The 128 original anchors were retained and 1920 additions came from fixed physical candidates
36 passing QC. This whole-NAM genotype selection includes the future target genotypes and is
37 transductive. It does not use target yield records, but may affect representation of rare or family-
38 confined variants. Subsequent task transforms use the source core and do not repeat the global 5%
39 MAF filter. NAM and MAGIC cached genotypes were checked against the same source PLINK
40 BED files for the 2048 markers. Their internal coding correspondence does not establish a
41 nucleotide-level interpretation of every numerical allele label or an absence of biological identity
42 concerns.

43 An early 128-marker fingerprint screen flagged candidate groups and near-matching pairs. Sparse-
44 marker matches are not proof of whole-genome identity, seed contamination or phenotype leakage.
45 The predefined MAGIC five-ID group was assigned entirely to A. A comprehensive new dense-
46 genome duplicate-exclusion experiment was not performed. IRRI's dense identity rule is separate and
47 must not be generalized to NAM/MAGIC. Any frozen-prediction candidate-domain sensitivity is
48 identified as a change of scoring population rather than a correction of demonstrated tuning leakage.

49 The Spindel archive of International Rice Research Institute (IRRI) trials is independent of Wei
50 NAM/MAGIC as a source archive. It is shared with TrG2P Rice299 and several training-set studies,
51 which use different screening and prediction designs. The material-season yield records are equal

52 means of at least two finite numeric Kg_ha observations with distinct archive REP identifiers; they
53 do not reproduce the source study's adjusted analysis. Agreement with a released yield record
54 verifies value correspondence, not the upstream field phenotype construction.

55 Original IRRI qualification grouped records by GHID, year and season; season codes 1 and 2 denote
56 dry and wet seasons. Any repeated REP entry within a material-season made the entire material-
57 season ineligible, whether PLOTNO was identical or differed; no value-based duplicate resolution
58 was applied. After whitespace stripping, blank strings, NA, N/A, NaN, nan, NULL, null, '.', '-' and '?'
59 were treated as missing. Remaining values had to be float-parsable and finite; no range, outlier, zero
60 or sign filter was used. Any unrecognized non-numeric value would have stopped processing; none
61 occurred. A qualified material-season required at least two distinct valid REP records and a finite
62 equal-weight arithmetic mean, without missing-replicate imputation. The prior metadata audit found
63 no empty REP/PLOTNO fields or conflicting GHID ownership of a REP/PLOTNO slot. These
64 identifiers do not by themselves establish biological independence. The resulting means were neither
65 block-adjusted values nor BLUEs.

66 **Table S1. NAM family origins and numbers of materials with finite released grain-yield records.**

Family	Non-HHZ parent	Parental genetic group	All materials	SH finite yield	HZ finite yield
N1	Kasalath	Aus	2192	2135	2088
N2	Basmati	Basmati	1693	1587	1561
N3	IAC25	Tropical japonica	1088	1052	1051
N4	Tebonnet	Tropical japonica	1218	1194	1176
N5	Wuyungeng7	Temperate japonica	861	846	827
N6	Xiushui134	Temperate japonica	869	845	845
N7	Nipponbare	Temperate japonica	1064	1034	994
N8	Chugeng37	Temperate japonica	930	913	886
N9	9311	Indica	896	891	885
N10	Zhongzao35	Indica	971	957	939
N11	IJ124	Intermediate	1289	1232	1160
N12	Changmi446	Indica	781	776	766
N13	Xiangwanxian12	Indica	838	818	836
N14	Fengyuezhan	Indica	659	656	655
N15	Guinongzhan	Indica	670	656	667
Total	15 non-HHZ parents	6 groups	16019	15592	15336

67 All families share the indica cultivar Huanghuazhan (HHZ). Non-HHZ parental groups follow Wei et al. (2024), Figure
68 1, and describe the parental accessions rather than ancestry proportions of recombinant lines. Family–parent links in Data
69 S4 were confirmed by exact material-ID agreement with the corresponding figshare family-map headers (Chen 2026).
70 Counts refer to the complete released family before calibration/scoring splits; finite yield records were counted in
71 Shanghai and Hangzhou columns K and Z of Data S4. They are not scoring-set sizes or independent field replicates.

72 **S2. NAM family tasks and information accounting**

73 SH scoring IDs follow the archived deterministic ID-hash split recorded in the author's NAM plan
74 file, `nam_plan.json`. The public release (Li 2026) lists the scoring and calibration identities of every
75 scored task in `resources/study.json.gz`; the source, core and pseudo-target allocations, needed for
76 refitting but not for re-scoring, remain in the plan file, which is available from the author on
77 reasonable request. Each family S contains $\text{floor}(N/2)$ eligible IDs. Adaptation candidates supply
78 nested C25/C100 draws, taken as the first 25 and 100 lines of a pseudo-random order fixed in

79 advance by a hash of target family, draw and line identifier, at HZ from HZ-eligible lines outside the
80 SH scoring sets; original arms, the calibrated-Gaussian completion and five draws share S. The
81 retained plan fixes source/core/pseudo-target/C/S order and settings; no split was regenerated. HZ
82 uses the SH S intersection with HZ eligibility. Restricted SH reranks the stored SH predictions on
83 that intersection and is not an independent site.

84 For each outer target family, two source-family pseudo-targets have 350 yield records each: 100
85 pseudo-calibration and 250 pseudo-validation. The other 12 source families supply a 1200-record
86 core. Core plus both pseudo-target blocks forms common1900. Extra source yield records expand
87 this to 1975 or 2000. Final allocations are source2000+C0, source1975+C25 and source1900+C100.
88 The 700 pseudo-target yield records are already inside the source information budget, not an
89 additional charge. Target-only100 uses those 700 development yield records and C100, totaling 800;
90 only C100 supplies its final fitting responses.

91 In each inner fold, all 350 records of the current pseudo-target family were removed from the
92 relevant source bank. Source-only arms fitted 1650 source records; calibrated arms fitted 1625
93 source records plus pseudo-C25 or 1550 source records plus pseudo-C100, and each was evaluated
94 on the 250 withheld validation records. The other pseudo-target family's 350 records remained
95 eligible as source records. Target-only100 fitted only pseudo-C100. After parameter selection,
96 development records reentered the final source banks. Across the five paired configurations, the
97 target calibration set, pseudo-target family pair and source/core samples varied, while the scoring set
98 remained fixed. Tuning ties within $10^{-12} \times \max(1, |\text{best mean } U|)$ favored larger λ , then smaller η
99 where applicable, then bandwidth multipliers in the order 1, 0.5, 2.

100 Source means and population standard deviations transform genotypes; zero-variance columns are
101 omitted. The linear kernel is $\frac{ZZ'}{p}$; here p is the number of retained markers, rather than an allele
102 frequency. Gaussian distances are squared standardized Euclidean distances divided by p, with $b^2 = c$
103 $\times$ median non-diagonal source-core distance, $c = 0.5/1/2$. Normalized target-mass-weighted loss, an
104 unpenalized intercept and raw-scale yield records are common. The ridge grid is $\lambda =$

0.0001/0.001/0.01/0.1; the target-mass grid is $\eta = 0.05/0.20/0.50$. Each arm chooses its own parameters by equal mean pseudo-family top-10% utility. No actual target-S yield record enters tuning. The same family's yield records may serve as source records in another outer task. Five draws are paired within each family and all compared arms use the same scoring IDs in that task. Source-only cells have 2000 permitted source yield records; calibrated cells replace 100 source yield records with the current target-family C100. A family may supply source records in another outer task. Role isolation therefore applies to each configuration and does not create independent validation populations or a globally inaccessible phenotype file.

The original family-domain descriptor table is retained as workbook sheet AF38 and original_csv/AF38.csv in Additional file 3. It contains all 45 records with N_S , selected q , recorded-yield mean, five-draw source-ridge r and ΔL in U/r . Added tables provide budget-fraction bounds, absolute arms and descriptive baseline associations. For SH, $\text{floor}(N_{\text{eligible}}/2) = N_S$ implies N_{eligible} in $\{2N_S, 2N_S+1\}$; both bounds are retained rather than guessing parity. Phenotypic SD, measured heritability and donor-level genomic mechanisms were not newly estimated.

119 **S3. Loss normalization, regularization and variance conventions**

NAM/MAGIC use source-core genotype means and population standard deviations, omitting source-degenerate columns. The ridge kernel is ZZ'/p , where p is the retained marker count. The Gaussian kernel and normalized objective are defined in the main Methods. The intercept is unpenalized. Each source row in target-mass-weighted ridge receives $(1-\eta)/n_{\text{source}}$ and each target row η/k , so total mass is one. At source1900+C100, $\eta = 0.05$ gives uniform final weights; pseudo-target fitting compositions can nevertheless yield different selected settings.

For the equal-weight NAM linear arms, let Z be the source-core standardized marker matrix before dividing by $\sqrt{p}$ and $X = \frac{Z}{\sqrt{p}}$, so $K = XX' = \frac{ZZ'}{p}$. Here p is the number of retained markers, b is the linear coefficient vector rather than the scalar Gaussian bandwidth, and n is the number of training rows. The fitted objective is mean squared residual plus λ times the kernel norm,

equivalently $\lambda \|b\|^2$ for $f = Xb$. Under the specific model $g \sim N(0, \sigma_g^2 K)$ and independent equal-
variance residuals $e \sim N(0, \sigma_e^2 I)$, with the same unpenalized intercept:

$$\lambda = \frac{\sigma_e^2}{n \sigma_g^2}$$

Thus $n = 2000$ gives $n\lambda = 0.2/2/20/200$ at the four grid values. This is a model-implied variance-ratio
convention, not measured heritability; cross-validation-selected λ values are not empirical genetic-
variance estimates. For raw-Z coefficients $\beta = \frac{b}{\sqrt{p}}$, the same penalty is $\lambda p \|\beta\|^2$. Domain-mass
weights, a differently scaled kernel, family/environment residual covariance or a different nuisance
design require corresponding model assumptions and do not inherit the simple equal-weight
interpretation. The finite λ grid does not identify a biological mechanism.

Matching numeric λ across two banks holds the penalty convention fixed while changing the training
genotypes, responses and kernel spectrum. Grid maxima observed on S describe the evaluated
settings; deployment settings were selected using source-family pseudo-target data.

S4. MAGIC partition and original model choices

The MAGIC population derives from (Huanghuazhan $\times$ Kasalath) $\times$ (Nipponbare $\times$ IAC25),
followed by selfing (Wei et al. 2024). The MAGIC crosswalk has 2402 identities; 2194 jointly finite
SH/HZ records form A/S halves of 1097 each. A previously audited five-ID genotype component
stays wholly in A. Component membership affects the split rather than deleting individuals. The
fixed source pseudo-targets are N12 and N10, each with 100 adaptation plus 250 validation yield
records; the 1200-row core comes from 13 other NAM families. The source banks and budget
accounting match the 1900/1975/2000 design above. MAGIC's active source-core marker count is
2044 from the historical 2048-marker panel.

Original selected settings are retained: ridge0 $\lambda = 0.01$ at both sites; RKHS0 $\lambda = 0.001$ and squared-
bandwidth multiplier 2; weighted25 $\lambda = 0.1$, $\eta = 0.05$ at SH and $\lambda = 0.01$, $\eta = 0.05$ at HZ;
weighted100 $\lambda = 0.001$, $\eta = 0.50$ at SH and $\lambda = 0.01$, $\eta = 0.05$ at HZ; pool100 $\lambda = 0.1/0.01$ at SH/HZ;

154 pooledRKHS100 $\lambda = 0.0001$, multiplier 1; target-only100 $\lambda = 0.001/0.0001$. They describe original
 155 arm-specific choices rather than a scoring-set oracle. Archive arm keys are retained in canonical
 156 tables: RKHS0 is source Gaussian0 and pooledRKHS100 is pooled Gaussian100; the NAM
 157 completion uses RKHS100 for calibrated Gaussian100. These are display-name mappings, not
 158 additional fitted models.

159 The original analysis used 240 source-development and 70 final fits. Source-only predictions are
 160 unchanged across C draws. Later matched- λ calculations retained $\eta = 0.50$ at SH and 0.05 at HZ,
 161 generating 120 logical records from 68 unique predictors, of which 17 were reused and 51 newly
 162 fitted. At HZ, weighted and pooled final row weights coincide at every matched λ ; this does not
 163 imply identical pseudo-development fits.

164 The final candidate-domain sensitivity made zero new fits. Its existing common-allele proxy uses
 165 1491 markers; the original S mean = 0.7173784120707227 and population SD =
 166 0.05895971125830521 give threshold ≥ 0.8942575458456383 and 25 IDs. The proxy is not ancestry
 167 or contamination proof. Removing those IDs only from S changes N/q from 1097/110 to 1072/108. It
 168 leaves A, C, training, preprocessing, parameters and predictions intact. Existing 190 domain rows
 169 and 230 contrasts/decompositions are indexed; original and reduced-domain means are never
 170 interchanged in the primary results. No winsorization, extra cutoff, alternative selection fraction or
 171 new inference was added.

172 Table 3 reports the original Shanghai allocation interaction to three decimal places (-0.004 g/plant);
 173 full precision (-0.004396969894549585) is provided in Additional file 2, sheet AF06, and in the
 174 canonical CSV in Additional file 3.

175 The five calibration draws overlap and share the same S, source core and tuning design. Their
 176 observed minima and maxima describe only those allocations; they are not confidence intervals or
 177 five independent population replicates. No new interval is supplied for MAGIC, so the sign and
 178 range of a recorded contrast do not establish precision for a wider breeding population.

179 S5. IRRI qualification and original deployment

180 The verified archive starts with 108024 dosage markers. Source call rate ≥ 0.90 , source minor-allele
181 frequency ≥ 0.01 , incompatible-allele exclusion and material call rate ≥ 0.90 on the qualified panel
182 retain 64208 markers and 201 source IDs. Final preprocessing removes three source-zero-variance
183 features, yielding 64205. Qualification uses source genotypes for marker filtering; target eligibility
184 uses the fixed panel. The high-concordance identity rule required at least 10000 comparable sites
185 spanning at least 10 chromosomes, with at least 100 joint calls on each and $\geq 99.9\%$ concordance.

186 No exclusion component resulted. The archived A1303 and A1203 records are not treated as aliases.

187 Source observations are 2009 DS/WS and 2010 DS/WS, with 110/159/161/196 material-season
188 records, totaling 626 from 201 IDs. The new-entry target cohort is defined from first appearance in
189 2011 DS; genotype qualification retains 121 IDs. A frozen deterministic ID-hash rule chose the
190 original C25 once. The same IDs support four separate seasonal calibrations. Scoring N is
191 96/96/94/96 and $q = 10$ at $\alpha = 0.10$. Each fit keeps all 626 historical observations and only the
192 current season's 25 yield records. Across seasons these are up to 100 material-season yield records
193 from the same 25 IDs.

194 For the original IRRI analysis and the post-hoc fixed policy, source pseudo-target 2009 WS uses a
195 110-ID/110-observation source bank and 24 scoring IDs after 25 pseudo-C IDs; 2010 WS uses 161
196 source IDs/430 observations and 15 scoring IDs after 25 pseudo-C IDs. Equal mean Pearson r over
197 the two scoring sets selects $\lambda = 0.01$ for ridge (mean $r = 0.268$) and $\lambda = 1$, $\gamma = 2$ for Gaussian (mean r
198 $= 0.415$). Twenty-four development fits cover three ridge and nine Gaussian candidates in each of
199 two batches. Real target S is excluded from this choice; boundary settings remain the finite-grid
200 selection rather than a global optimum or a mechanistic result.

201 Dosages use 0/2 with 255 for missingness. Source-unique genotypes define means, allele-frequency
202 scaling $\sqrt{2p(1-p)}$, linear-kernel trace normalization and the median source-pair squared distance.
203 Seasonal replicate rows reuse their material's transformed genotype. Here p denotes the source allele
204 frequency. The Gaussian is:

205
$$\exp\left(-\frac{\gamma d^2}{\text{median}_{\text{source pairs}}(d^2)}\right)$$

206 Full training-season indicator columns are jointly unpenalized; there is no extra global intercept
 207 column. Target-season dummy coefficients in calibrated fits use C only. Source-only unseen-season
 208 indicators are zero, which leaves an arbitrary common prediction offset that cannot affect within-
 209 season ranking.

210 Target-similarity weighting uses the C-genotype centroid and the fixed median source-pair distance.
 211 Source weights $\exp(-d^2/\text{scale})$ are clipped to $[0.5, 2]$, then normalized to source-observation mean
 212 one; target rows have weight one. This is not the target-mass weighting used in NAM/MAGIC.
 213 Nuisance and kernel coefficients are jointly estimated.

214 The original analysis comprised 24 development fits, two reusable source-only final fits and 12
 215 calibrated final fits, for 38 fits. Four seasons follow the same target cohort and are correlated follow-
 216 ups. Only original 2011 DS pool25 – source Gaussian0 retains the original primary role. Its
 217 favorable outcome and the adverse 2012 WS pool25 – source ridge0 check are both retained.

218 **S6. Post-hoc four-arm completion and IRRI policy sensitivity**

219 The NAM completion retained source/core/C/S roles, markers, transforms, objective and arm-
 220 specific source-family U selection. It reused 450 original prediction vectors and 6000 inner records
 221 after checking that inputs, settings and outputs matched in order. The RKHS100 arm added 3600
 222 inner plus 150 final scientific fits, totaling 3750; 18 preselected technical replays were separate,
 223 totaling 3768 new physical fits. All contrasts, domains and adverse results were retained.

224 Technical refits used $\text{SH/HZ} \times \text{N1/N8/N15} \times \text{r0} \times \text{ridge0/RKHS0/pool100}$. Complete archived
 225 prediction vectors, ranking and selected IDs met the fixed tolerance checks. Separately implemented
 226 model-to-forward-prediction reconstruction also agreed within the recorded tolerances. These checks
 227 establish numerical and provenance agreement, not byte identity of newly refitted model archives.
 228 Historical scaler arrays lacked a separate fit-time fingerprint; no retrospective digest is claimed.

229 Historical NAM code read site-wide valid phenotypes, but its fitting and tuning used the permitted
 230 task indices and scoring followed parameter selection. Inspection of the code found no use of scoring
 231 yield records in fitting or tuning, and later code also restricted each task to its permitted records.

232 The fixed 20 C25 sets are deterministic draws from the retrospective intersection of 119 four-season-
 233 complete target IDs. Eligible seasonal counts are 121/121/119/121, so each scoring complement has
 234 96/96/94/96 IDs and $q = 10$. B1169 and M1480 remain in S for their eligible seasons. The source
 235 bank retains 626 material-season observations from 201 source IDs; calibrated policies add 25
 236 current-season observations. Historical 24 source-development fits are parameter provenance inside
 237 those 626 yield records, not 24 additional yield records.

238 Fixed IRRI policy retained linear $\lambda = 0.01$ and Gaussian $\lambda = 1, \gamma = 2$, requiring 242 final fits and zero
 239 new inner fits. The adaptive policy evaluated five arm-specific grids in two historical wet-season
 240 batches: each arm used $\lambda = 0.01/0.1/1$; both Gaussian arms also crossed $\gamma = 0.5/1/2$, giving 27
 241 candidates per batch and 54 inner fits. Calibrated candidates fitted pseudo-C25 and validated on
 242 pseudo-S24/15 with $q = 3/2$.

243 Adaptive selection maximized the equal arithmetic mean of the two pseudo-target U values. Values
 244 within $10^{-12} \times \max(1, |\text{best mean U}|)$ were tied. The predefined ordering preferred larger λ , then γ
 245 closest to one, then smaller γ . Locked settings were $\lambda = 0.1$ for source ridge; $\lambda = 1$ for pooled and
 246 similarity-weighted ridge; and $\lambda = 1, \gamma = 1$ for both Gaussian arms. These settings were used
 247 unchanged across all 20 outer configurations and four seasons.

248 Adaptive pooled and similarity-weighted ridge tied at $\lambda = 0.1$ and 1 on both fold U values. Source
 249 Gaussian had six tied candidates, all three λ crossed with $\gamma = 1$ or 2; pooled Gaussian had three:
 250 $(0.1, 2)$, $(1, 1)$ and $(1, 2)$. The rule selected $(1, 1)$ in both Gaussian arms. A selected-set objective can
 251 form such plateaus, but the observed ties do not certify that any setting is uniquely optimal, establish
 252 a universal q threshold or isolate a causal regularization mechanism.

253 Each policy generated 400 arm-by-configuration-by-season cells, for 800 in total. Each fixed source-
 254 only model can be reused across configurations, while changing S still changes candidates and U.

255 Configurations overlap and their spread combines changes in calibration and scoring complements.
 256 All five arms, all seasons, positive and negative cases and literal zeros are retained. The historical
 257 allocation remains separate from C01–C20; it is not configuration 21.
 258 The full NAM completion contains 81 negative absolute-U cells, 50 negative defined-r cells and five
 259 cells with negative r but positive U. The 800 IRRI arm cells have defined r throughout and include
 260 45 negative absolute-U cells. These are distinct endpoint descriptions, not counts of failed fits or
 261 interchangeable signs of model validity.

262 **S7. Post-hoc summaries and frozen-prediction sensitivities**

263 Interval and correlation summaries were added after earlier results had been inspected, using frozen
 264 scored tables, contrasts and the original NAM bootstrap-index matrix. The subsequent direct D,
 265 absolute-arm, decision-threshold/rank-correlation and present family-dispersion summaries are post
 266 hoc. Frozen-table calculations did not access raw phenotypes; prediction rescoring verified archived
 267 predictions before accessing scoring phenotypes, without refitting or retuning. The fixed- λ r
 268 extension additionally used the exact archived B1 prediction vectors after the analysis settings were
 269 fixed and the scoring code was checked separately, without new fits or tuning.

270 NAM has three scoring views $\times$ 15 families $\times$ five draws $\times$ four arms. Each family contributes the
 271 mean of five paired draw differences; summaries weight 15 families equally. Contrasts are $\Delta L =$
 272 pooled ridge100 – source ridge0; $\Delta G =$ calibrated Gaussian100 – source Gaussian0; $K0 =$ source
 273 Gaussian0 – source ridge0; $K100 =$ calibrated Gaussian100 – pooled ridge100; $I = \Delta G - \Delta L = K100$
 274 $- K0$; and $D =$ pooled ridge100 – source Gaussian0 $= \Delta L - K0 = \Delta G - K100 = (\Delta L + \Delta G - K0 -$
 275 $K100)/2$. D is the difference between the average target-record and kernel increments of the tested
 276 complete strategies, including arm-specific reoptimization.

277 The original 10000×15 index array is reused in its original $N1, \dots, N15$ order, without generating
 278 random indices. Each row samples paired family means with replacement and averages them.
 279 Linear-interpolated 2.5th/97.5th percentiles give conditional pointwise descriptions. The t sensitivity
 280 uses the sample SD with 14 degrees of freedom and $t_{0.975,14} = 2.144786687916927$. Delete-one-

281 family summaries average the remaining 14 frozen family values after deleting each family once;
 282 their range is a sensitivity descriptor, not a confidence interval or a refitting analysis.
 283 The archived two-arm fixed- λ predictions were additionally rescored for r at all four penalties in all
 284 three NAM scoring views. The 150 original tasks supplied 1200 prediction vectors and 1800 arm-by-
 285 view scores; all 5400 numerical checks against the original 900 paired U rows reproduced exactly
 286 before new results were accepted. Five draw-specific r differences were averaged within family,
 287 followed by equal averaging over all 15 families. All 12 grid summaries and their positive/exact-
 288 zero/negative family counts are reported descriptively, without new intervals or p-values (Additional
 289 files 2 and 3).
 290 All r summaries average task-specific correlations arithmetically, without Fisher transformation or a
 291 pooled candidate-level correlation. Existing contrast, D and absolute-arm families contain 30, six and
 292 24 pointwise intervals. A further result-informed sensitivity treats only ΔL , ΔG and D at SH/HZ as
 293 six comparisons: original-bootstrap quantiles use probabilities 0.0041666667/0.9958333333, and t_{14}
 294 intervals use critical value 3.068779245476067. These are nominal 95% familywise Bonferroni
 295 sensitivities for the selected six, conditional on the same family-resampling or t working
 296 assumptions; they do not cover every analysis or establish independent-family sampling.
 297 Selected-six r sensitivity results. All four target-record bootstrap intervals excluded zero after the
 298 selected-six adjustment. The t_{14} sensitivity excluded zero for SH ΔL , SH ΔG and HZ ΔG , while HZ
 299 ΔL and both D intervals spanned zero; the adjusted bootstrap interval also spanned zero for HZ D.
 300 These results distinguish positive mean strategy effects from method-dependent interval support.
 301 Six selected predictive-ability contrasts: Bonferroni sensitivity at nominal familywise 95%.

Site	Contrast	Mean Δr	Bootstrap interval	t_{14} interval
SH	ΔL	+0.049417	[+0.021812, +0.075688]	[+0.016133, +0.082702]
SH	ΔG	+0.058706	[+0.024739, +0.103960]	[+0.010672, +0.106741]
SH	D	+0.038258	[+0.006558, +0.075992]	[-0.003980, +0.080497]

Site	Contrast	Mean Δr	Bootstrap interval	t ₁₄ interval
HZ	ΔL	+0.027430	[+0.003957, +0.057236]	[−0.004578, +0.059439]
HZ	ΔG	+0.044952	[+0.019434, +0.081152]	[+0.007554, +0.082349]
HZ	D	+0.022722	[−0.001079, +0.047260]	[−0.006962, +0.052406]

302 Relative-r descriptors retain both $100 \times \text{mean}(\Delta r_f)/\text{mean}(r_{\text{baseline},f})$ and $\text{mean}[100 \times$
 303 $\Delta r_f/r_{\text{baseline},f}]$. They are ratios of means and means of ratios, respectively. SH ridge/Gaussian
 304 values are 26.4%/29.6% versus 32.5%/20.7%; HZ values are 16.8%/26.8% versus 13.0%/−7.4%.
 305 Negative baseline families number one at SH and two at HZ for each model. Thus the HZ Gaussian
 306 summaries even differ in sign. No family is removed to improve a percentage, zero/near-zero rules
 307 remain explicit and no ratio interval is supplied. These are not genetic-accuracy or realized-gain
 308 percentages; absolute r and Δr are the primary summaries, and the relative value quoted in the main
 309 text is the ratio of means.

310 Descriptive cross-site baseline–gain check. Across 15 frozen family means, source-ridge SH baseline
 311 r versus HZ ΔL_r was −0.400759, and HZ baseline versus SH ΔL_r was −0.624226; baseline r
 312 across sites was 0.929638. Matched Gaussian-baseline values versus the other site's ΔG_r were
 313 −0.656363/−0.769337. With B and C denoting source and calibrated family r , $\text{Cov}(B_{\text{SH}}, C_{\text{HZ}} -$
 314 $B_{\text{HZ}}) = \text{Cov}(B_{\text{SH}}, C_{\text{HZ}}) - \text{Cov}(B_{\text{SH}}, B_{\text{HZ}})$. This identity alone does not establish bias.
 315 Under a latent-score model with signal–error orthogonality and independent cross-site errors, the
 316 covariance can reflect a cross-site baseline–gain relationship. Those assumptions were not
 317 established here; the sites reuse scoring materials and related family/source information, and score-
 318 error covariances were not estimated. The associations therefore describe these scores rather than a
 319 prospectively validated allocation rule.

320 Observed between-family dispersion. For each site and strategy pair, B and C are the source and
 321 calibrated five-draw family means of r for all 15 families; SD uses denominator 14. All six SD ratios
 322 below one show a smaller observed spread of family scores. Oldham's descriptor is $\text{corr}(C - B,$
 323 $(C+B)/2)$, and $\text{Cov}(C - B, (C+B)/2) = [\text{Var}(C) - \text{Var}(B)]/2$. It describes the same variance difference

on another scale, rather than an independent confirmation of a mechanism. Interpreting it as a relationship between latent initial predictability and latent gain requires assumptions about signal–error orthogonality and error variance (Oldham 1962; Chiolero et al. 2013). Equal error variances were not established by using the same S or nearby r values. The table provides descriptive point values without new p-values, confidence intervals or an allocation-policy claim. D compares complete strategies with different model classes. Non-positive five-draw mean r counts fell from 1/15 to 0/15 at SH for both models and from 2/15 to 1/15 for ridge and to 0/15 for Gaussian at HZ. These are marginal descriptors over all families; uncertainty around near-zero scores was not estimated.

Site / contrast	SD(B) → SD(C)	SD ratio C/B	r(B,C)	Oldham r
SH / ΔL	0.117 → 0.095	0.810	0.943	−0.539
SH / ΔG	0.135 → 0.098	0.722	0.914	−0.634
SH / D	0.135 → 0.095	0.704	0.952	−0.759
HZ / ΔL	0.130 → 0.116	0.892	0.952	−0.352
HZ / ΔG	0.138 → 0.109	0.787	0.955	−0.631
HZ / D	0.138 → 0.116	0.837	0.972	−0.608

IRRI compares arms within the same policy, configuration and season. Six contrasts are retained: pool25 – RKHS0, pool25 – ridge0, RKHS0 – ridge0, weighted25 – pool25, pooled Gaussian25 – pool25, and pooled Gaussian25 – RKHS0. Same-S adaptive-minus-fixed arm differences and differences between policy contrasts remain separate. Seasonal means, medians, sample SDs, minima/maxima and positive/exact-zero/negative counts describe the 20 overlapping configurations. Signs use literal >0, <0 and ==0, without numerical binning. U/r agreement distinguishes same nonzero sign, opposite nonzero sign, both zero and single-endpoint zero.

For fixed pool25 – ridge0, U/r same-sign counts are 13/20, 7/20, 10/20 and 11/20, totaling 41/80; adaptive totals 38/80. Fixed pool25 – RKHS0 totals 57/80, while adaptive totals 41/80 with two U-only zeros. The comparison denominator is therefore always specified. No binomial, independent-cell t test or cell bootstrap is used. Source-only policy comparisons retain partitions, grids,

transformations and final source yield records while changing the selection rule; calibrated-arm comparisons also include deployment-specific reoptimization.

IRRI top-5% comparator sensitivity. Values are pooled ridge25 minus the indicated source-only strategy, in kg/ha. Every row summarizes the same 20 overlapping configurations within that policy and season; +/0/- denotes literal positive/exact-zero/negative counts. Means and all counts for both comparators are shown together. Medians and complete-precision fields are retained in IRRI_top5 and its canonical CSV. These rows transcribe existing frozen results without new scoring.

Policy / season	ΔU vs source ridge	+/0/-	ΔU vs source Gaussian	+/0/-
fixed / 2011DS	+245.9	16/1/3	+57.9	12/0/8
fixed / 2011WS	+227.4	15/0/5	-347.4	4/0/16
fixed / 2012DS	+231.6	15/0/5	+101.6	14/0/6
fixed / 2012WS	+212.6	16/0/4	-290.4	1/0/19
adaptive / 2011DS	+167.0	12/2/6	+65.5	13/4/3
adaptive / 2011WS	+145.9	12/0/8	-39.4	7/3/10
adaptive / 2012DS	+93.6	10/2/8	-115.8	2/4/14
adaptive / 2012WS	+56.0	11/1/8	-144.1	3/3/14

Selection-fraction and Spearman additions. The main frozen-prediction extension retains 900 NAM scoring cells, including restricted SH, and 800 IRRI cells, evaluated at $\alpha = 0.05/0.10/0.20$. Its 5100 metric rows and 188000 selected-ID rows were written without refitting, retuning, new predictions, bootstrap draws or new intervals. Original 10%-selection metrics were replayed by parsed-float exact equality, including 53820 original selected rank/ID/prediction/truth rows. Spearman is computed once per whole-S cell with average tied ranks and repeated across α only for table alignment. All 1700 whole-S cells were defined in this completed run; the source still retains explicit undefined handling.

The original weighted ridge100 – source Gaussian0 comparison is separately bound to its historical prediction provenance and source tables. Its companion contains 225 whole-S cells, 675 metric rows, 675 paired-draw contrasts, 135 family rows and 27 descriptive summaries. The original 450 arm-cell

362 records at $\alpha = 0.10$ and selected-ID order/N/q were replayed; 2308 legacy value checks had
 363 maximum absolute difference 4.44×10^{-16} under the fixed absolute/relative 10^{-10} tolerance. Original
 364 AF03/AF04 values and interval references remain unchanged; no new interval was estimated for
 365 other fractions or correlations. The companion is not a fifth four-strategy arm.

366 The distinction between full-cohort association and selection enrichment holds within individual
 367 families. For HZ N7 at 5%, original-primary $\Delta U = +0.39563$ g/plant despite $\Delta r = -0.00548$ and $\Delta p =$
 368 -0.01382 . For HZ N12 at 20%, $\Delta U = -0.58919$ despite $\Delta r = +0.00066$ and $\Delta p = +0.01834$. These are
 369 retained counterexamples, not new tests.

370 IRRI top10 replacement accounting retains all 160 fixed/adaptive configuration-season pairs, with
 371 ordered selected lists, common/entered/exited IDs and recorded-yield contributions. The maximum
 372 absolute arithmetic residual was approximately 9.1×10^{-13} kg/ha. In source ridge, the U rule uniquely
 373 selected $\lambda = 0.1$ rather than 0.01 on the same historical validation design, and r fell in all 80
 374 overlapping scoring cells. Source Gaussian retained $\lambda = 1$ but changed $\gamma = 2$ to $\gamma = 1$ through a six-
 375 way exact U plateau and the tie rule; r rose in 6/14/20/19 configurations by season (59/80). These are
 376 finite-design consequences of the declared selection rules, with overlapping scoring cells rather than
 377 independent trials.

378 Additional file 2 contains seven main-extension tables (M_metrics, M_NAM_fam, M_NAM_arms,
 379 M_NAM_ctr_fam, M_NAM_contrasts, M_IRRI_arms, M_IRRI_contrasts), five original-primary
 380 tables (P_metrics, P_Spearman, P_paired, P_families, P_summary), NAM_r_multiplicity,
 381 IRRI_top5, NAM_family_dispersion, and four fixed- λ r tables (Grid_r_arms, Grid_r_pairs,
 382 Grid_r_families, Grid_r_summary). Additional file 3 contains their canonical CSVs, retaining earlier
 383 scientific CSV content unchanged and mapping the original filenames to AF02–AF38 in
 384 MANIFEST.json. The dispersion and top-5% tables reorganize frozen results without additional
 385 fitting, prediction generation, raw-phenotype scoring, p-values or resampling intervals. The archived
 386 public reproduction package produces selected-ID order digests from its frozen predictions (Li 2026;
 387 Section S11).

388 S8. Earlier descriptive sensitivities and selected-set accounting

389 Shanghai all/same- λ /different- λ task means are +0.82/+0.39/+1.09 g/plant, with 75/29/46 tasks and
390 15/13/15 families. Positive tasks are 48/19/29. N10/N15 are absent from the same- λ subset. Equal
391 means of represented-family subset means are +0.39 g/plant for same λ and +0.87 g/plant for
392 different λ . The task-weighted and family-equal same- λ values differ in the unrounded evidence even
393 though both round to +0.39 g/plant. Hangzhou paired task means are +0.33/+0.51/+0.10 g/plant, with
394 75/41/34 tasks and 15/15/14 families; positive tasks are 43/28/15. N5 is absent from the different- λ
395 subset; same λ has two to five tasks per family. Corresponding family-equal subset means are
396 +0.49/+0.04 g/plant. Full precision is preserved in the CSV/JSON evidence.

397 The original selected- λ summaries retain six rows and 90 family-representation rows as AF11/AF12
398 in the workbook and original CSV members. Full task identities remain with the analysis code.

399 These are post-hoc subsets selected on chosen hyperparameters, nonrandom and uneven across
400 families. They are not a matched- λ intervention or replacement for the original main estimate.

401 Matching numeric λ does not hold training yield records or effective kernel spectra fixed.

402 In the original IRRI 2011 DS allocation, pooling shared six selected IDs with either source reference
403 and had four unique selected IDs against each. Common IDs cancel from ΔU . For pool25 – RKHS0,
404 unique-set means were 5719.2 and 5140.5 kg/ha; division of their unrounded sum difference by $q =$

405 10 gives +231.5 kg/ha. For pool25 – ridge0, the corresponding means were 5203.8 and 4565.9,
406 giving +255.1. The four unique IDs on each side do not define four independent biological pairs.

407 Complete identities and recorded contributions remain in Additional files 2 and 3 (original tables
408 AF13-AF14).

409 All 382 prediction values differ between source Gaussian0 and pooled Gaussian25. In 2011 DS/2011
410 WS/2012 DS/2012 WS order, prediction-pair Pearson correlations are 0.938/0.957/0.943/0.975 and
411 Spearman correlations are 0.954/0.979/0.955/0.984. The top-10 sets intersect 10/10 in every season,
412 while their head order and full-rank order differ. The complete 40 head records and 382 prediction-
413 pair records preserve individual ranks and values. After removing each season's mean shift,

414 maximum absolute differences are 2.1/2.4/1.3/1.3 kg/ha; full precision is retained in the descriptive
415 table.

416 The primary-season correlation with recorded grain yield changed from 0.270 for source Gaussian0
417 to 0.392 for pooled Gaussian25 while U remained 217.2 kg/ha. In 2012 WS, pooled linear ridge had
418 higher correlation than ridge (0.513 versus 0.441) but lower U. Correlation and recorded yield in the
419 selected fraction therefore answer different questions. Raw prediction-level differences include the
420 source-only unseen-season offset convention. Common translation or positive scaling alone cannot
421 explain changed correlation with recorded grain yield. The descriptors establish changed predictions
422 and rankings, not the biological mechanism responsible, and do not prove a strong-regularization
423 explanation from boundary $\lambda = 1$. Nor do identical selected sets imply that target yield records were
424 ignored.

425 The original NAM pooling-check bootstrap intervals remain +0.819 [0.296, 1.332] g/plant at SH and
426 +0.325 [-0.094, 0.797] at HZ. Their repetition in the four-arm matrix does not promote them to a
427 new primary. The original weighted ridge100 – source Gaussian0 primary and the new diagonal D =
428 pool100 – RKHS0 are distinct complete-strategy comparisons; the original checking contrast is
429 pool100 – ridge0.

430 Existing task-normalized enrichment expresses $100U/\text{mean}(S)$, and model differences are
431 percentage-point changes $100(U_A - U_B)/\text{mean}(S)$. Normalization is performed within the actual
432 task before the design-specific aggregation. It is distinct from a percentage relative to another arm's
433 U. The full 154 existing summaries remain in Additional file 2, sheet AF21, with the canonical CSV
434 in Additional file 3. Restricted SH at $\lambda = 0.1$ illustrates the difference: mean original-unit $\Delta U =$
435 -0.017946 g/plant, but task-normalized mean difference = $+0.063658$ percentage points. Neither
436 scale substitutes for the other, and no cross-design pooled effect or ratio interval is claimed.

437 **S9. Historical development branches**

438 Earlier development evaluated explicit interaction models, support weighting, denser additive
439 baselines, model-selection policies and a fixed neural architecture. The historical table retains mixed

440 and unfavorable outcomes. Their training budgets, tuning policies, candidate domains and sometimes
 441 marker panels differ from the current calibrated deployments; they are not additional matched
 442 benchmarks or independent confirmation. In the table, RMSE is root-mean-square error, MSE is
 443 mean squared error and Δ denotes the difference between the two named models.

Historical branch	Saved SH / HZ result	Interpretation
128-marker explicit interactions	All-interaction B – additive A Δ RMSE +0.295177 / +0.682704	Unfavorable in that grid; different panel and full-family test domain.
Support-weighted interactions	C – uniform U Δ RMSE +0.314037 / –0.001369; SH14/15 worse	No consistent benefit in this fixed diagnostic.
Out-of-fold error decomposition	B – A macro MSE reduction: 76.34% /81.69% from mean-bias ² term	Oracle-centering arithmetic, not a deployable calibrator or causal fraction.
Dense additive + I128	Δ RMSE –0.151814 /–0.024289; Δ r +0.006547 /–0.000603	Limited increments; HZ adverse r retained.
Model-selection policy pilot	Gu – Rr Δ U –0.639203 /+0.172574; original conditional intervals [–1.288318, –0.029370] /[–0.206339, +0.541815]	Both validation grouping and selection criterion change.
Fixed neural architecture	N4 – N1 Δ U +0.114 /–0.179; N4 – ridge –1.316 /–1.219	622 formal fits; small-range classification is not equivalence.

444 The source-only model-selection pilot simultaneously changed validation grouping and the selection
 445 criterion. Its adverse SH Gu – Rr U difference limits a universal recommendation for grouped U
 446 tuning, but does not identify the isolated effect of U. Its selected gradient-boosted tree policy had U
 447 = 6.123332/3.657912 g/plant at SH/HZ on complete held-out families. It trained on all eligible yield
 448 records from the other 12 families, unlike the current half-family S and 2000-record scenario. Those
 449 values cannot be ranked directly against current pooled ridge100.

450 The neural branch completed 622 formal fits (167 reused, 455 new) within 630 historical attempts
 451 and stopped on 4 October 2026. Its fixed 2048→64→16→4 architecture, three-seed prediction

452 averaging, auxiliary traits and supervision weights differ from the current five-draw calibrated
 453 strategies. Its frozen N4 – N1 classification was small_range_difference, not equivalence or general
 454 evidence against multi-trait neural prediction. Negative N4-versus-ridge outcomes remain disclosed.

455 **S10. Analysis chronology and evidence roles**

456 Internal freezing documents chronology; it is not public preregistration. The register separates
 457 original comparisons from later result-informed extensions. Negative, positive and zero outcomes
 458 follow the same inclusion rules.

Analysis	Chronology / role	Retained interpretation
Original NAM/MAGIC	Internally frozen after earlier development; weighted100 – RKHS0 primary; pooling checks	Complete strategy comparisons. No public preregistration.
Original IRRI allocation	2011DS pool25 – RKHS0 primary; same-cohort seasonal follow-ups	One original C25 allocation; follow- ups correlated.
Stored sensitivities	Post-hoc fixed- λ , chosen- λ subsets, MAGIC matched- λ /domain, normalized descriptors and primary t intervals	Original primary and full-domain outcomes retained.
NAM four-arm completion	Post-hoc protocol before new calibrated-Gaussian results; 3750 scientific fits +18 technical	Same budget and per-arm source- family U tuning; five strategy contrasts.
IRRI fixed20	Post-hoc C25/S partition sensitivity; 242 final fits, zero new tuning	20 overlapping configurations; original allocation excluded.
IRRI adaptive20	Separate post-hoc policy layer; 54 inner +242 final fits; one locked setting per arm	Criterion and calibrated reoptimization change together.
Post-hoc companion summaries	After initial result inspection; frozen-table r/U intervals and paired descriptors	No fitting or tuning; later frozen- prediction rescoring separately reaccesses truth.

Analysis	Chronology / role	Retained interpretation
Direct strategy and threshold sensitivities	Later post-hoc D/absolute-arm summaries; 5%, 10%, 20% rescoring and Spearman	Frozen predictions unchanged; separate weighted-primary companion preserves the original 10% identity. Fixed IRRI keeps r-oriented settings.
Selected-six r sensitivity	Post-hoc, after inspection of earlier results; original 10000 index rows reused	Six chosen contrasts only; conditional, method-dependent interval support.
Fixed- λ r rescoring	Post-hoc, after earlier results; protocol locked before scoring 1200 archived prediction vectors	12 descriptive summaries; no new fits, intervals or tests.

459 **S11. Numerical verification and code availability**

460 Separately implemented scoring and table checks verified the retained outputs. The four-arm
461 completion used preselected old-arm replays and separately implemented forward reconstruction;
462 IRRI preserved its source/marker order and all stored scoring values against their historical identities.
463 The historical IRRI process did not record a separate genotype-content digest at fit time, so current
464 retained-input identity is not a newly created historical process fingerprint. These provenance
465 distinctions concern the computational record, not evidence that scoring yield records entered tuning.

466 Charts were generated programmatically from stored result tables using ReportLab, and PNG
467 versions were rasterized from vector PDFs using PDFium. Models were fitted and scored with
468 custom Python/NumPy implementations; the archived reproduction reference execution used Python
469 3.12 and NumPy 2.3.5. Bounded numerical and synthetic scoring checks on Windows and macOS,
470 preselected old-arm replays, and separately implemented model-to-prediction reconstruction support
471 the verified numerical interfaces. Agreement with established packages such as rrBLUP (Endelman
472 2011) was not separately benchmarked. The public frozen-prediction reproduction package (Li 2026)
473 is available at <https://github.com/yuriljx/rice-genomic-selection-validation/releases/tag/v1.0.0> and
474 archived as software version 1.0.0 at <https://doi.org/10.5281/zenodo.23438608>. It includes author-
475 generated frozen predictions, exact scoring identities, numerical model implementations, source
476 checks and one scoring entry. With a locally obtained copy of the Wei et al. (2024) Data S4

workbook and Dryad version 1 file RYT_plotdata_by_GHID_corrected_PUBLIC_ACCESS.zip, the entry reproduces 4940 frozen arm-by-view cells and 9240 metric rows for the original NAM/MAGIC/IRRI comparisons, NAM four-strategy and four-penalty analyses, and both IRRI policies over all 20 partitions. It does not refit or retune models, re-run the historical tuning scripts, reproduce every exploratory branch or MAGIC matched-penalty/proxy-domain extension, or reconstruct Word layouts. Its input adapters use only the scored records from the hash-checked source files and, for IRRI, average their replicate observations; they do not repeat the full historical qualification described in Section S1. Software/documentation are MIT-licensed; author-generated predictions/results are CC BY 4.0; source-derived identities retain their original attribution. The source phenotype files and raw genotypes are not redistributed.

These post-hoc analyses summarize stored results and rescore frozen predictions at decision thresholds and with rank correlation. The selected-six r sensitivity reuses the original 10000 bootstrap-index rows. The present dispersion descriptors use existing five-draw family means, and the top-5% comparator table transcribes the existing seasonal summaries. No new empirical model, dataset or observation was introduced, and original primary comparisons and signed outcomes are preserved. The later fixed- λ r extension reports every grid setting descriptively and introduces no new interval or test.

S12. Additional-file guide

Workbook sheet / canonical CSV	Contents	Data rows
AF02 / original_csv/AF02.csv	NAM arm and view means	18
AF03 / original_csv/AF03.csv	NAM family primary contrasts	15
AF04 / original_csv/AF04.csv	NAM pre-specified contrasts and intervals	20
AF05 / original_csv/AF05.csv	MAGIC original arm means	14
AF06 / original_csv/AF06.csv	MAGIC original contrasts	22
AF07 / original_csv/AF07.csv	MAGIC post-hoc matched-lambda contrasts	24

Workbook sheet / canonical CSV	Contents	Data rows
AF08 / original_csv/AF08.csv	MAGIC candidate-domain contrasts	46
AF09 / original_csv/AF09.csv	IRRI arm metrics	20
AF10 / original_csv/AF10.csv	IRRI contrasts	24
AF11 / original_csv/AF11.csv	NAM selected-lambda subset summaries	6
AF12 / original_csv/AF12.csv	NAM subset family representation	90
AF13 / original_csv/AF13.csv	IRRI primary selected-set accounting	2
AF14 / original_csv/AF14.csv	IRRI primary selection ledger	28
AF15 / original_csv/AF15.csv	IRRI Gaussian prediction descriptions	4
AF16 / original_csv/AF16.csv	IRRI Gaussian head ranks	40
AF17 / original_csv/AF17.csv	NAM post-hoc fixed-grid paired tasks	900
AF18 / original_csv/AF18.csv	NAM post-hoc fixed-grid family means	180
AF19 / original_csv/AF19.csv	NAM post-hoc fixed-grid summaries	12
AF20 / original_csv/AF20.csv	NAM post-hoc fixed-grid conditional intervals	12
AF21 / original_csv/AF21.csv	Design-specific relative selection effects	154
AF22 / original_csv/AF22.csv	NAM four-arm metrics	900
AF23 / original_csv/AF23.csv	NAM five paired strategy contrasts	1125
AF24 / original_csv/AF24.csv	NAM five-draw family means	225
AF25 / original_csv/AF25.csv	NAM equal-family descriptive summaries	15
AF26 / original_csv/AF26.csv	IRRI both-policy arm metrics	800
AF27 / original_csv/AF27.csv	IRRI both-policy within-partition contrasts	960
AF28 / original_csv/AF28.csv	IRRI same-S adaptive-minus-fixed arm differences	400
AF29 / original_csv/AF29.csv	IRRI same-S differences between policy contrasts	480

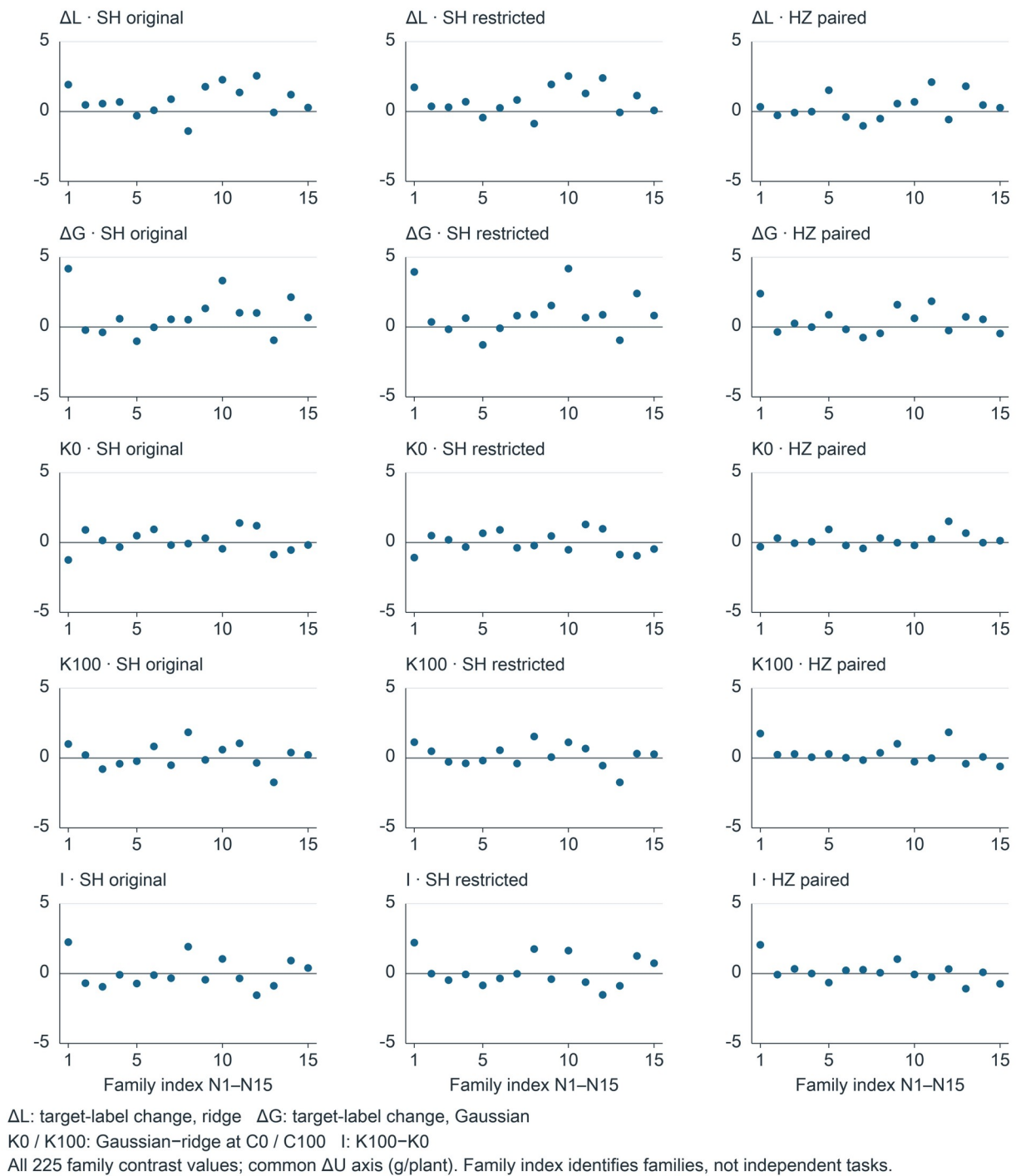
Workbook sheet / canonical CSV	Contents	Data rows
AF30 / original_csv/AF30.csv	IRRI complete seasonal descriptive summaries	192
AF31 / original_csv/AF31.csv	IRRI historical original-allocation metrics	20
AF32 / original_csv/AF32.csv	IRRI historical original-allocation contrasts	24
AF33 / original_csv/AF33.csv	NAM strategy contrasts and conditional intervals	30
AF34 / original_csv/AF34.csv	IRRI policy and season contrast summaries	48
AF35 / original_csv/AF35.csv	IRRI U and r sign agreement	48
AF36 / original_csv/AF36.csv	NAM matched-model relative r descriptors	6
AF37 / original_csv/AF37.csv	IRRI adaptive-minus-fixed policy summaries	20
AF38 / original_csv/AF38.csv	NAM family scoring domains and stored predictive summaries	45
N01 / new_tables/NAM_arm_family_means.csv	NAM four-arm five-draw family means	180
N02 / new_tables/NAM_arm_absolute_summary.csv	NAM absolute-arm summaries and conditional intervals	24
N03 / new_tables/NAM_directD_task_values.csv	NAM direct diagonal strategy task contrasts	450
N04 / new_tables/NAM_directD_family_values.csv	NAM direct diagonal family contrasts	90
N05 / new_tables/NAM_directD_summary.csv	NAM direct diagonal summaries and conditional intervals	6
N06 / new_tables/NAM_SH_label_budget_bounds.csv	NAM target-phenotyping fraction bounds	15
N07 / new_tables/NAM_AF38_descriptive_associations.csv	NAM descriptive baseline associations	6
N08 / new_tables/NAM_AF36_relative_ratio_checks.csv	NAM relative-r definition checks	6
N09 / new_tables/NAM_AF36_relative_family_values.csv	NAM complete family-specific relative-r values	90

Workbook sheet / canonical CSV	Contents	Data rows
M_metrics / new_tables/metrics.csv	Frozen-prediction endpoint metrics across selection fractions	5100
M_NAM_fam / new_tables/NAM_family_means.csv	NAM five-draw family means	540
M_NAM_arms / new_tables/NAM_arm_summary.csv	NAM equal-family absolute performance	108
M_NAM_ctr_fam / new_tables/NAM_contrast_family_means.csv	NAM paired family contrasts	2430
M_NAM_contrasts / new_tables/NAM_contrast_summary.csv	NAM contrasts across selection fractions	162
M_IRRI_arms / new_tables/IRRI_arm_summary.csv	IRRI arm performance across overlapping configurations	360
M_IRRI_contrasts / new_tables/IRRI_contrast_summary.csv	IRRI paired contrasts across selection fractions	144
P_metrics / new_tables/weighted_metrics.csv	Original NAM weighted-strategy companion metrics	675
P_Spearman / new_tables/weighted_Spearman.csv	Original NAM weighted-strategy Spearman companions	225
P_paired / new_tables/primary_paired_draws.csv	Original NAM primary paired draw contrasts	675
P_families / new_tables/primary_family_means.csv	Original NAM primary five-draw family contrasts	135
P_summary / new_tables/primary_summary.csv	Original NAM primary selection-fraction summaries	27
NAM_r_multiplicity / multiplicity/NAM_six_r_contrast_multiplicity_sensitivity.csv	Six selected r comparisons; Bonferroni interval sensitivity	6
NAM_family_dispersion / descriptors/NAM_family_dispersion_descriptors.csv	Observed family-level SD and change–midpoint descriptors	6

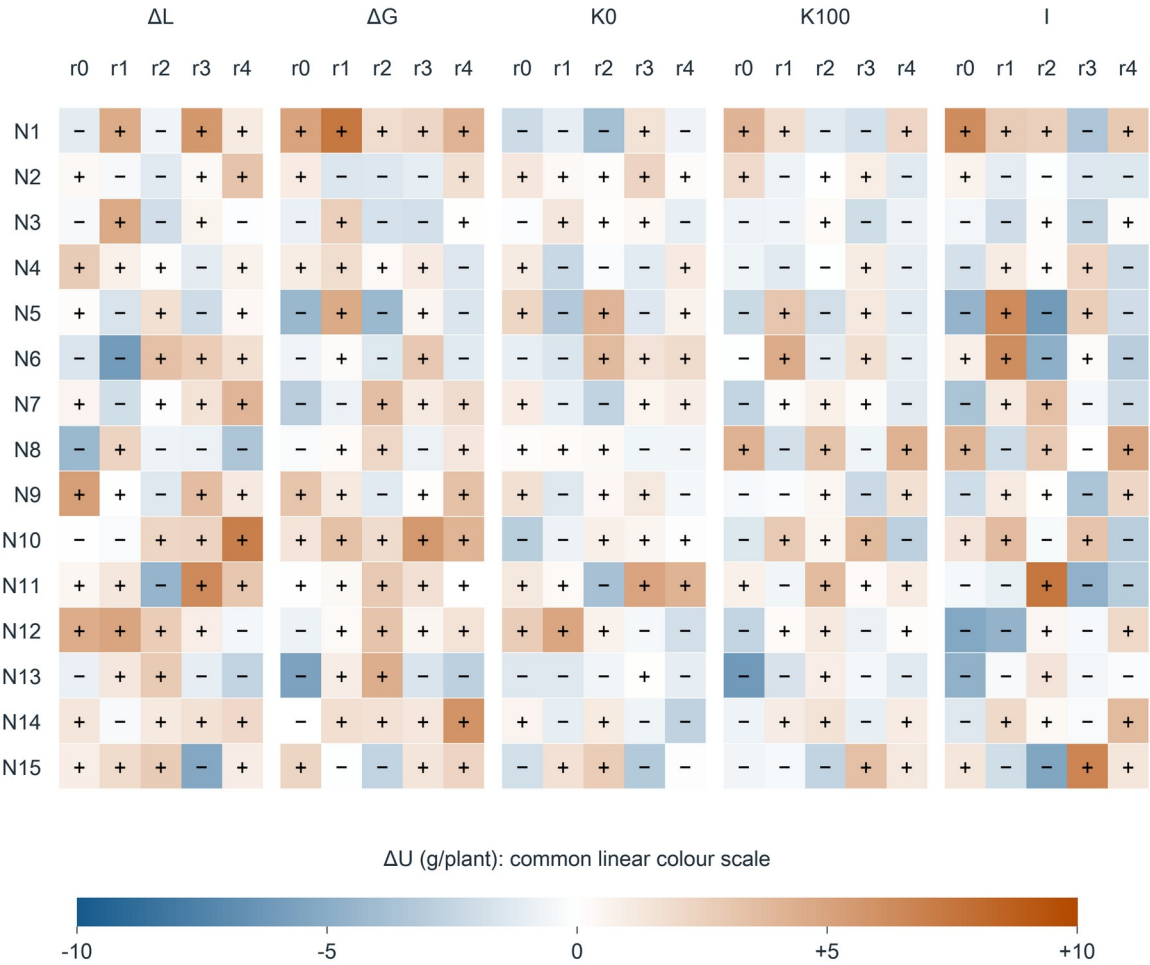
Workbook sheet / canonical CSV	Contents	Data rows
IRRI_top5 / descriptors/IRRI_top5_comparator_summary.csv	IRRI top-5% summaries for both source comparators	16
Grid_r_arms / fixed_grid/NAM_FIXED_GRID_R_ARM_TASKS.csv	Fixed- λ predictive ability: complete three-domain grid	1800
Grid_r_pairs / fixed_grid/NAM_FIXED_GRID_R_PAISED_TASKS.csv	Fixed- λ predictive ability: complete three-domain grid	900
Grid_r_families / fixed_grid/NAM_FIXED_GRID_R_FAMILY_MEAN_S.csv	Fixed- λ predictive ability: complete three-domain grid	180
Grid_r_summary / fixed_grid/NAM_FIXED_GRID_R_SUMMARY.csv	Fixed- λ predictive ability: complete three-domain grid	12

495 Additional file 1 is this complete supplement with Figures S1–S11. Additional file 2 is the readable
 496 result workbook. Additional file 3 contains the 37 original scientific CSVs, named AF02.csv–
 497 AF38.csv, together with the added result tables. MANIFEST.json records every CSV path, size and
 498 hash and maps each renamed original file to its former name. Workbook cells are text to preserve
 499 decimal strings and identities; numeric analyses should use the canonical CSVs. The table guide
 500 maps each table to a workbook sheet and ZIP member.

501 S13. Supplementary figures

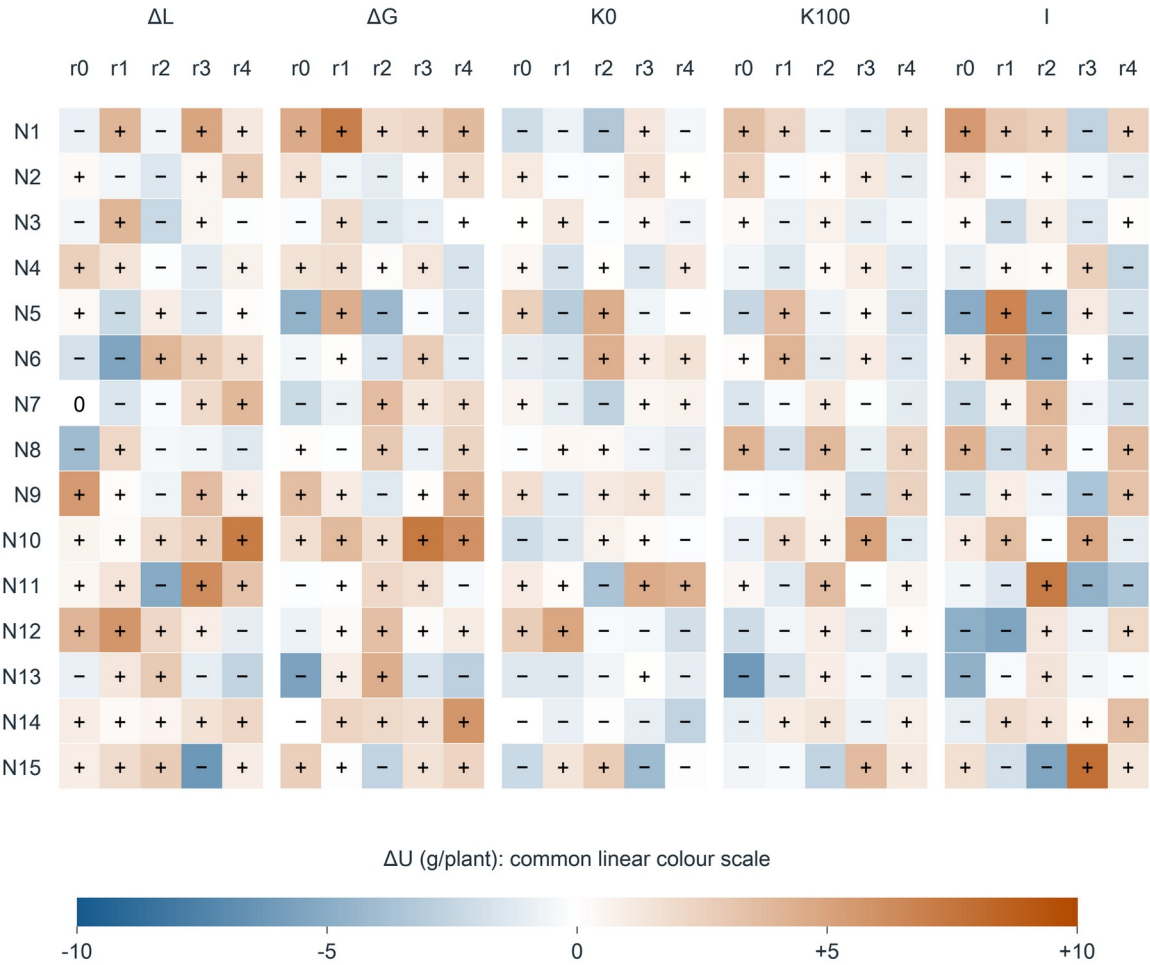


502
503 Figure S1. NAM family means. All five contrasts and three scoring views retain 15 five-draw family
504 means; units are g/plant.



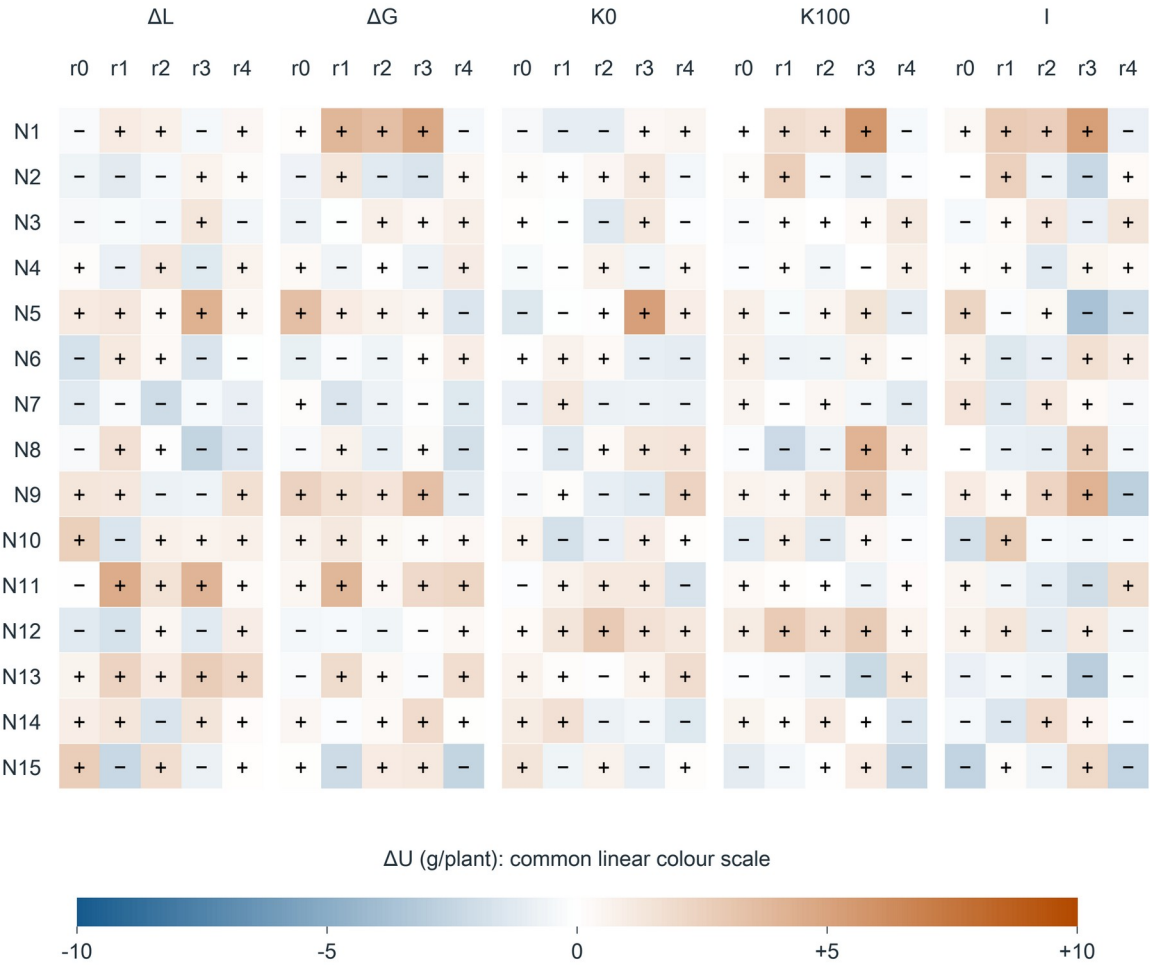
Natural family order N1–N15; stored draw identifiers r0–r4; – / + / 0 retain direction.

505
506 Figure S2. NAM draw-level contrasts. All 1125 values retain family, draw, scoring view and
507 contrast; colors show ΔU (g/plant) on a common scale. Page 1: SH original.



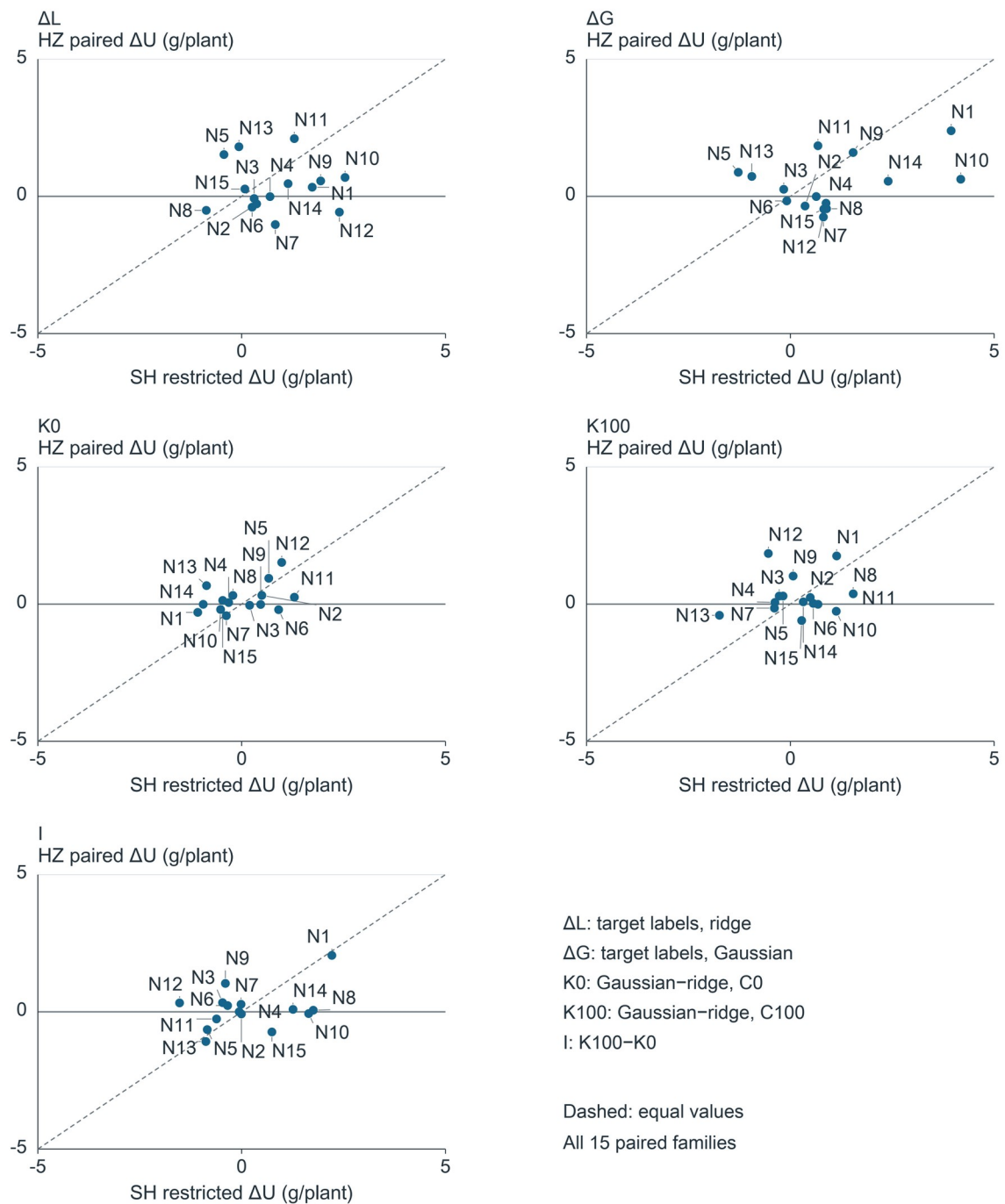
Natural family order N1–N15; stored draw identifiers r0–r4; – / + / 0 retain direction.

508
509 Figure S2. NAM draw-level contrasts. All 1125 values retain family, draw, scoring view and
510 contrast; colors show ΔU (g/plant) on a common scale. Page 2: restricted SH.



Natural family order N1–N15; stored draw identifiers r0–r4; - / + / 0 retain direction.

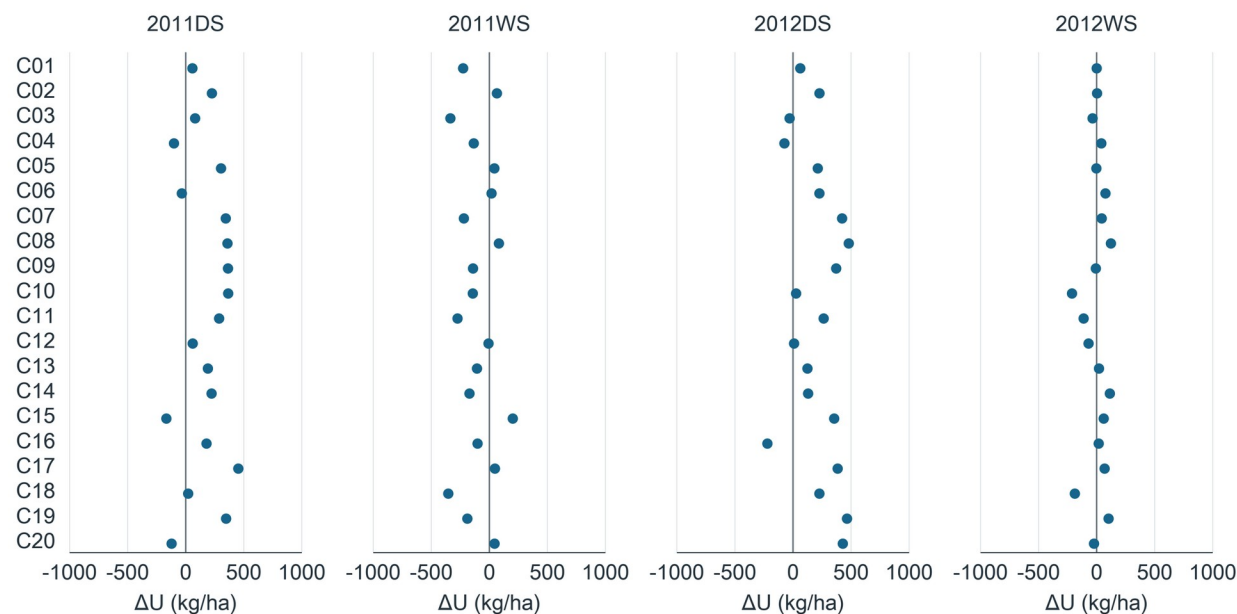
511
512 Figure S2. NAM draw-level contrasts. All 1125 values retain family, draw, scoring view and
513 contrast; colors show ΔU (g/plant) on a common scale. Page 3: HZ paired.



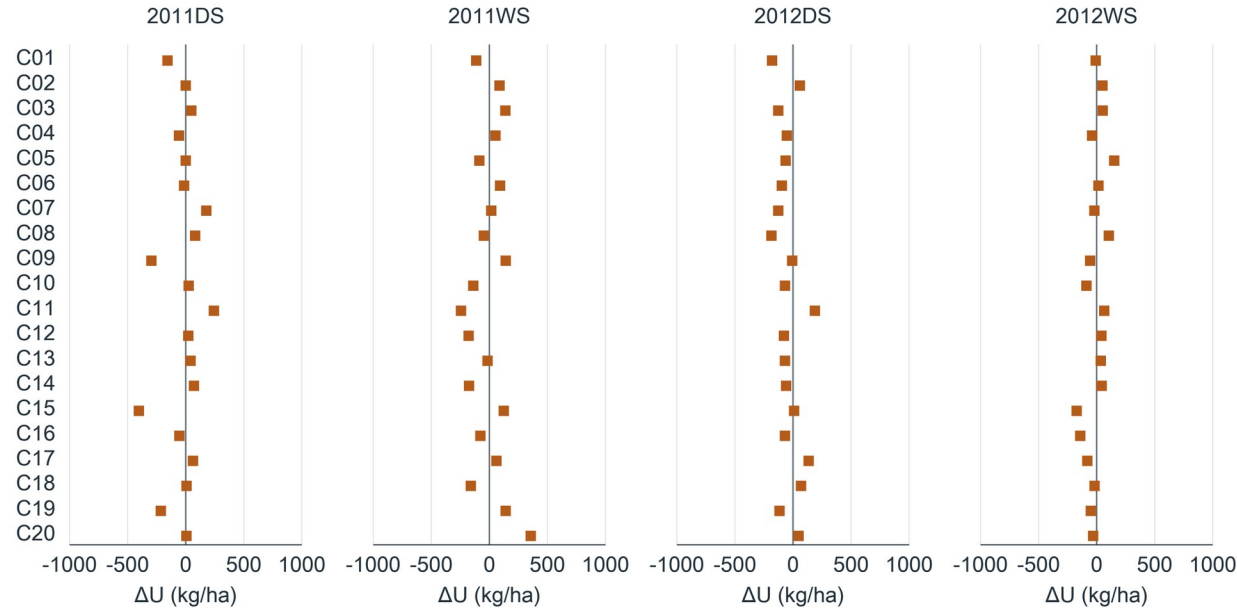
Common axes. Text moves near each point; source coordinates and family identities remain fixed.

514
 515 Figure S3. Paired scoring views. Family ΔU values use the same HZ-valid identities at restricted SH
 516 and HZ (g/plant); sites are not independent.

Fixed policy



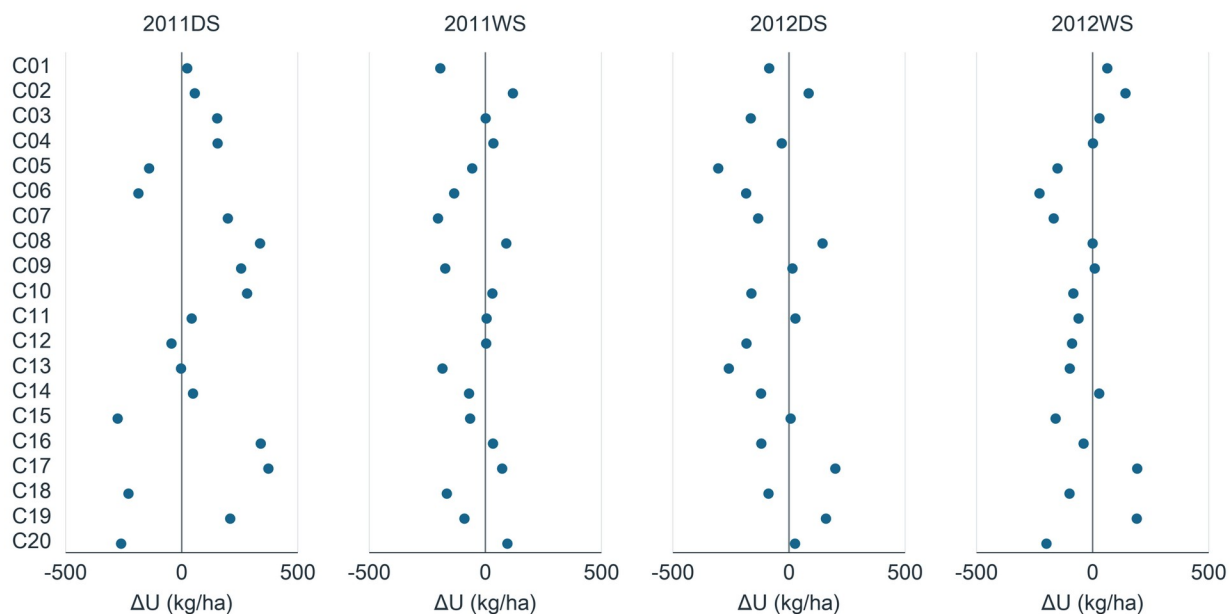
Adaptive policy



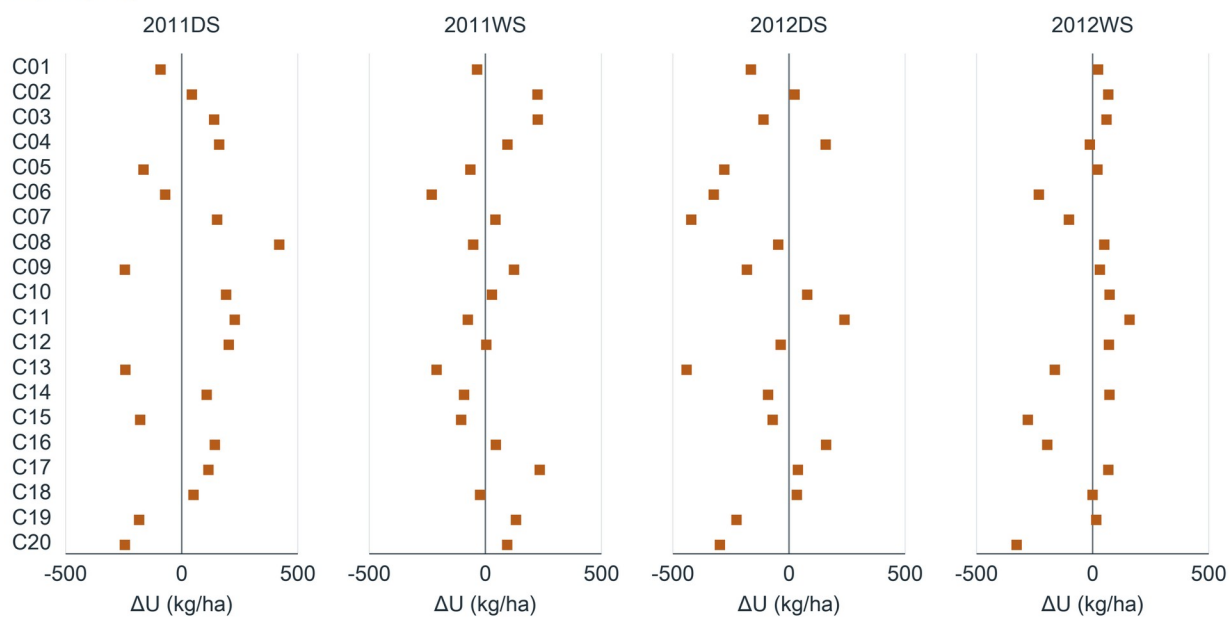
All C01–C20 identities follow the same top-to-bottom order in every panel; overlapping configurations.

517
518 Figure S4. IRRI pool25 – source Gaussian0. Both policies retain all 20 overlapping configurations
519 per season; ΔU is kg/ha, without confidence intervals.

Fixed policy

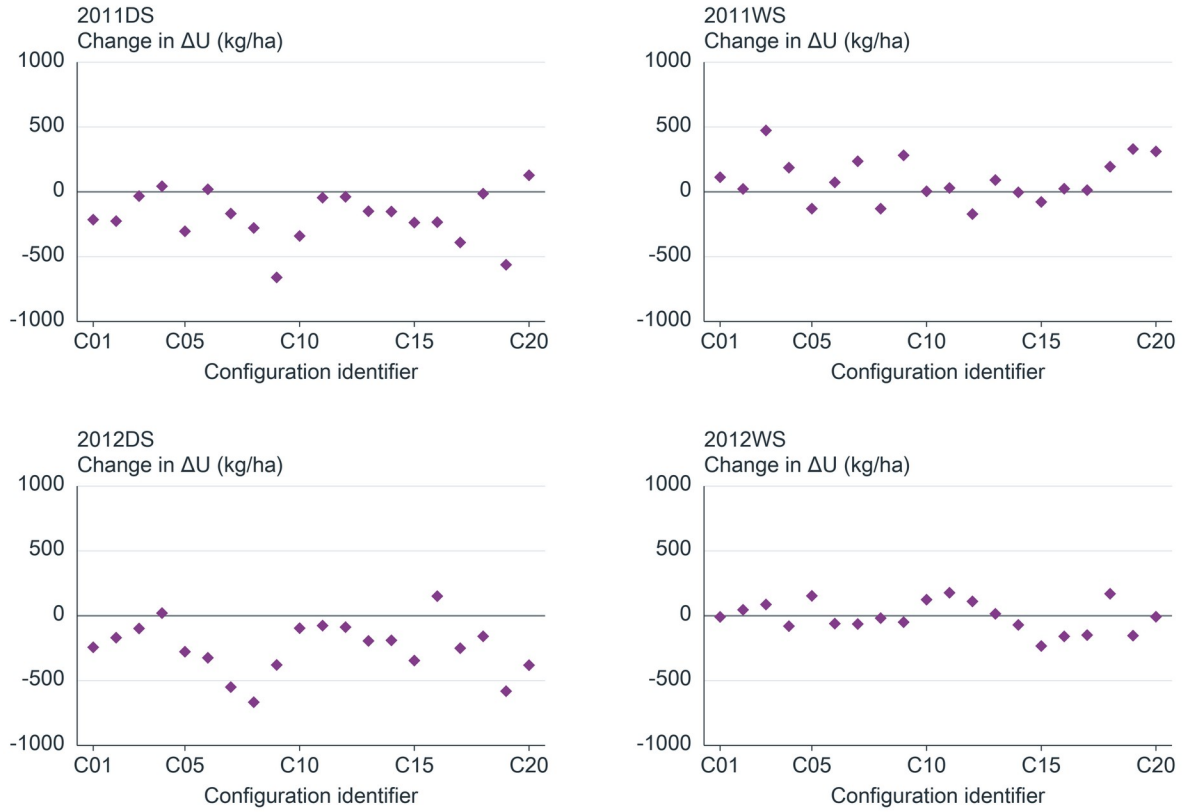


Adaptive policy



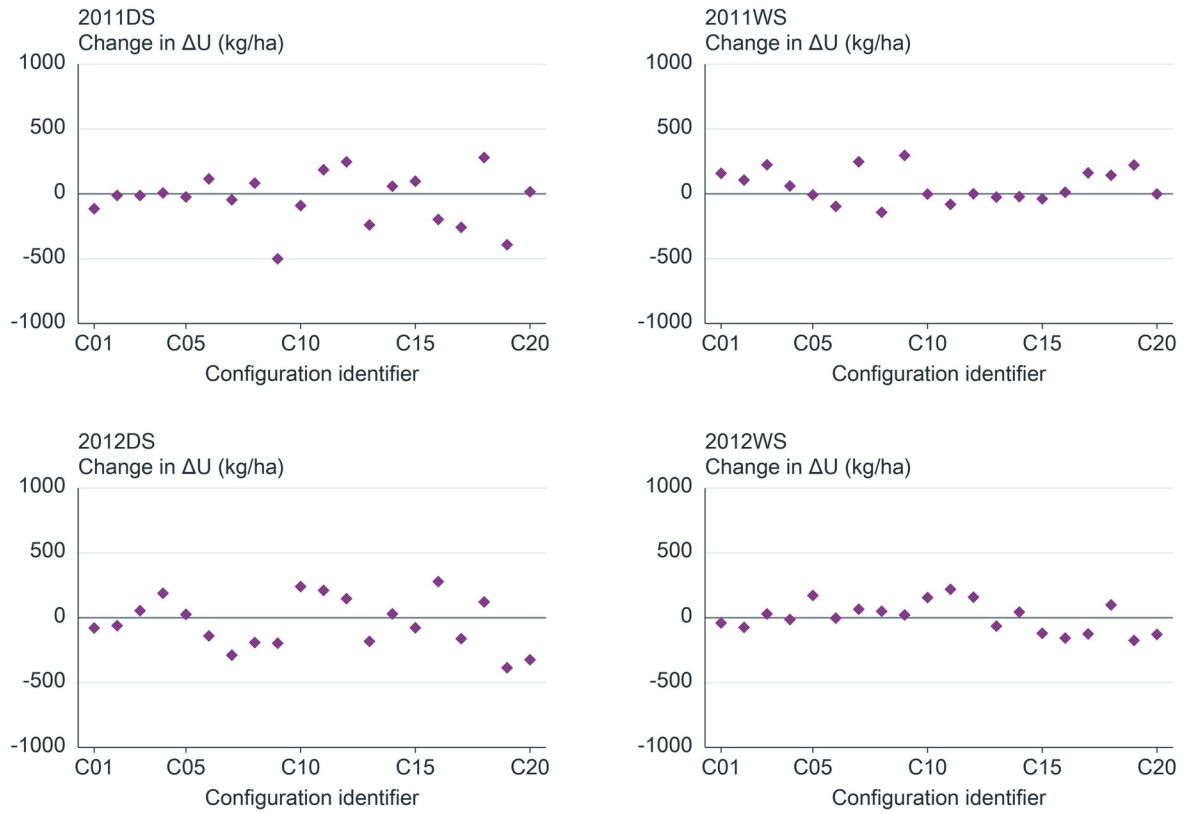
All C01–C20 identities follow the same top-to-bottom order in every panel; overlapping configurations.

520
 521 Figure S5. IRRI pool25 – source ridge0. Both policies retain all 20 overlapping configurations per
 522 season, including adverse ΔU values (kg/ha).



Same ordered S within a configuration. This compares complete policies, not an isolated tuning criterion.

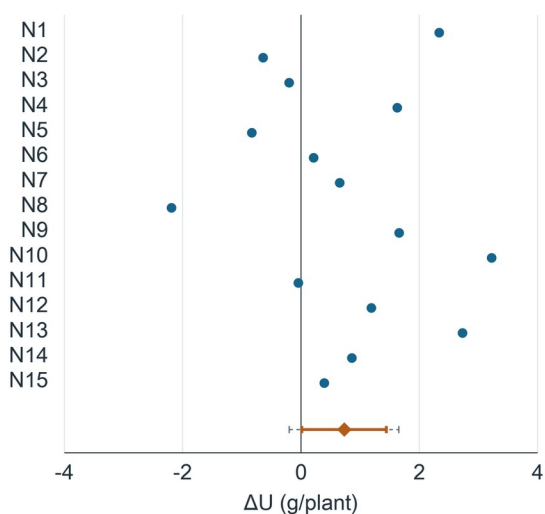
523
 524 Figure S6. IRR policy sensitivity versus Gaussian0. For each configuration and season, $\Delta\Delta U =$
 525 $[U(\text{pool25}) - U(\text{source Gaussian0})]$ under the adaptive policy minus the same contrast under the
 526 fixed policy, on the same S (kg/ha); the policy difference combines the tuning criterion and
 527 reoptimization.



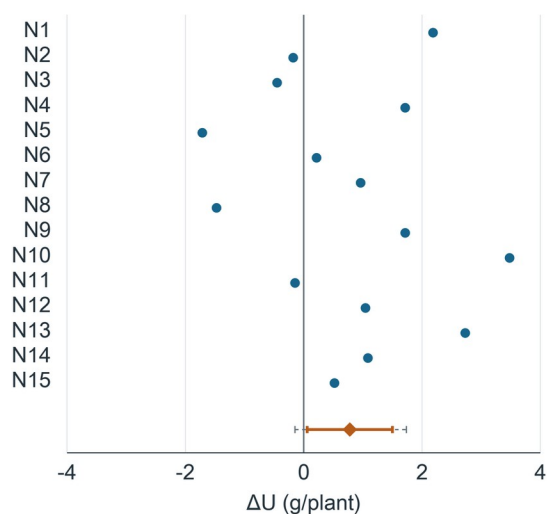
Same ordered S within a configuration. This compares complete policies, not an isolated tuning criterion.

528
 529 Figure S7. IRR policy sensitivity versus ridge0. For each configuration and season, $\Delta\Delta U =$
 530 $[U(\text{pool25}) - U(\text{source ridge0})]$ under the adaptive policy minus the same contrast under the fixed
 531 policy, on the same S (kg/ha); every season and configuration is retained.

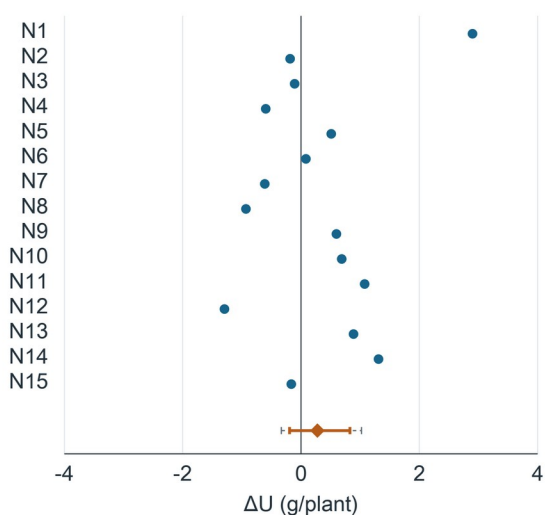
A Shanghai: original S



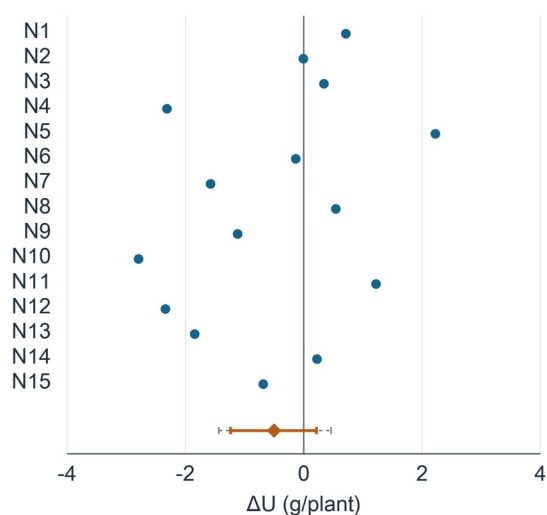
B Shanghai: restricted S



C Hangzhou: paired S



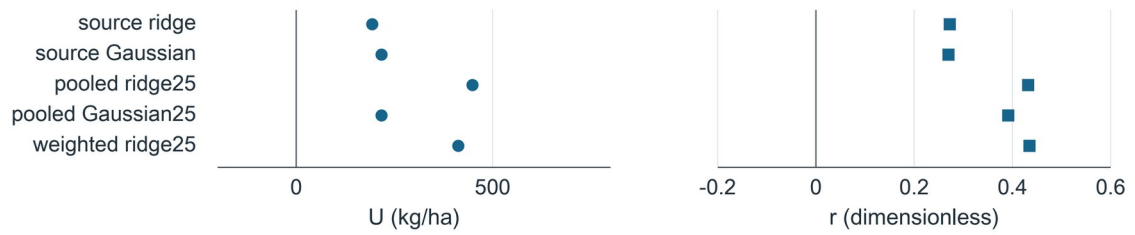
D Paired HZ minus SH



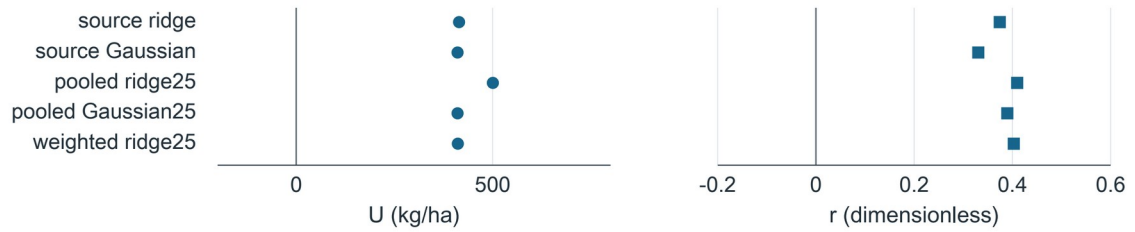
● Original family values ◆ Original mean / 95% solid / 99% dashed
weighted ridge100 – source Gaussian; original intervals retained without replacement.

532
533 Figure S8. Original NAM comparisons. Solid 95% and dashed 99% conditional bootstrap intervals
534 retain historical sensitivity displays, not joint coverage.

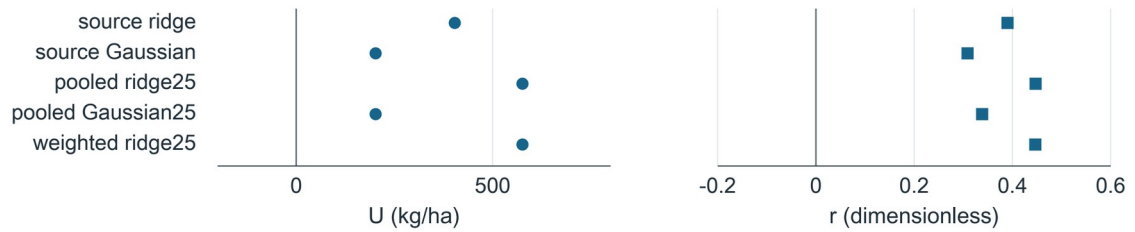
2011DS



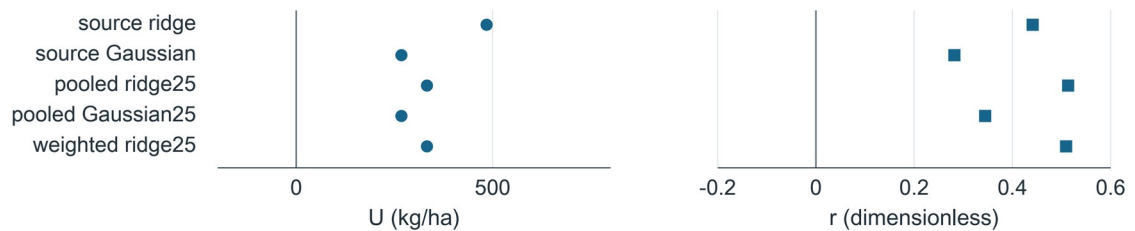
2011WS



2012DS



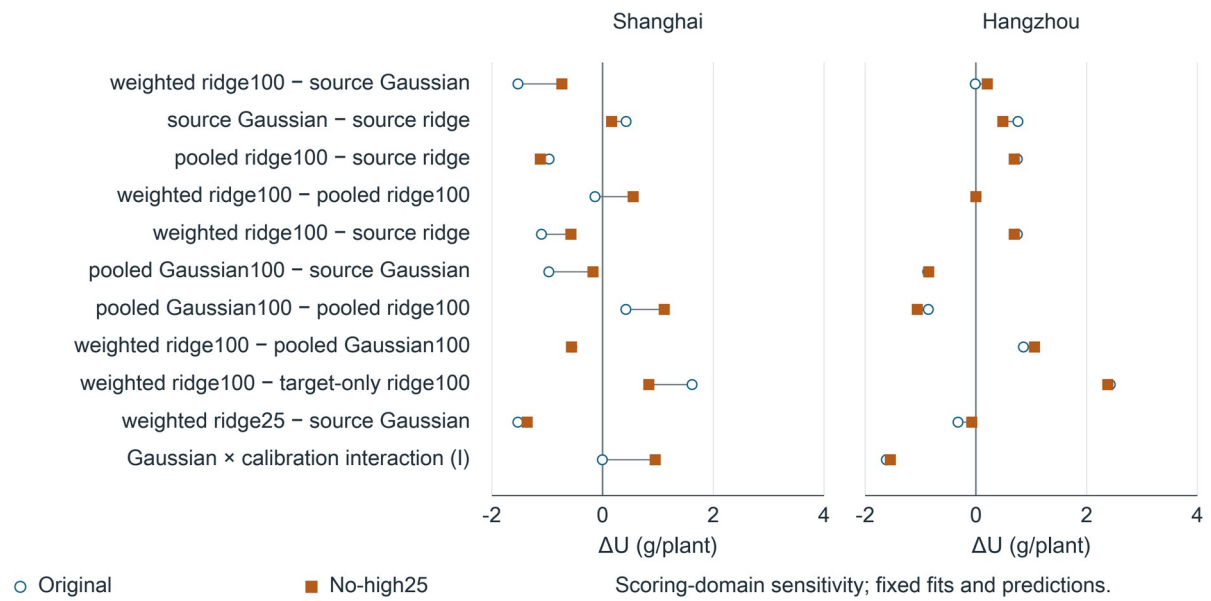
2012WS



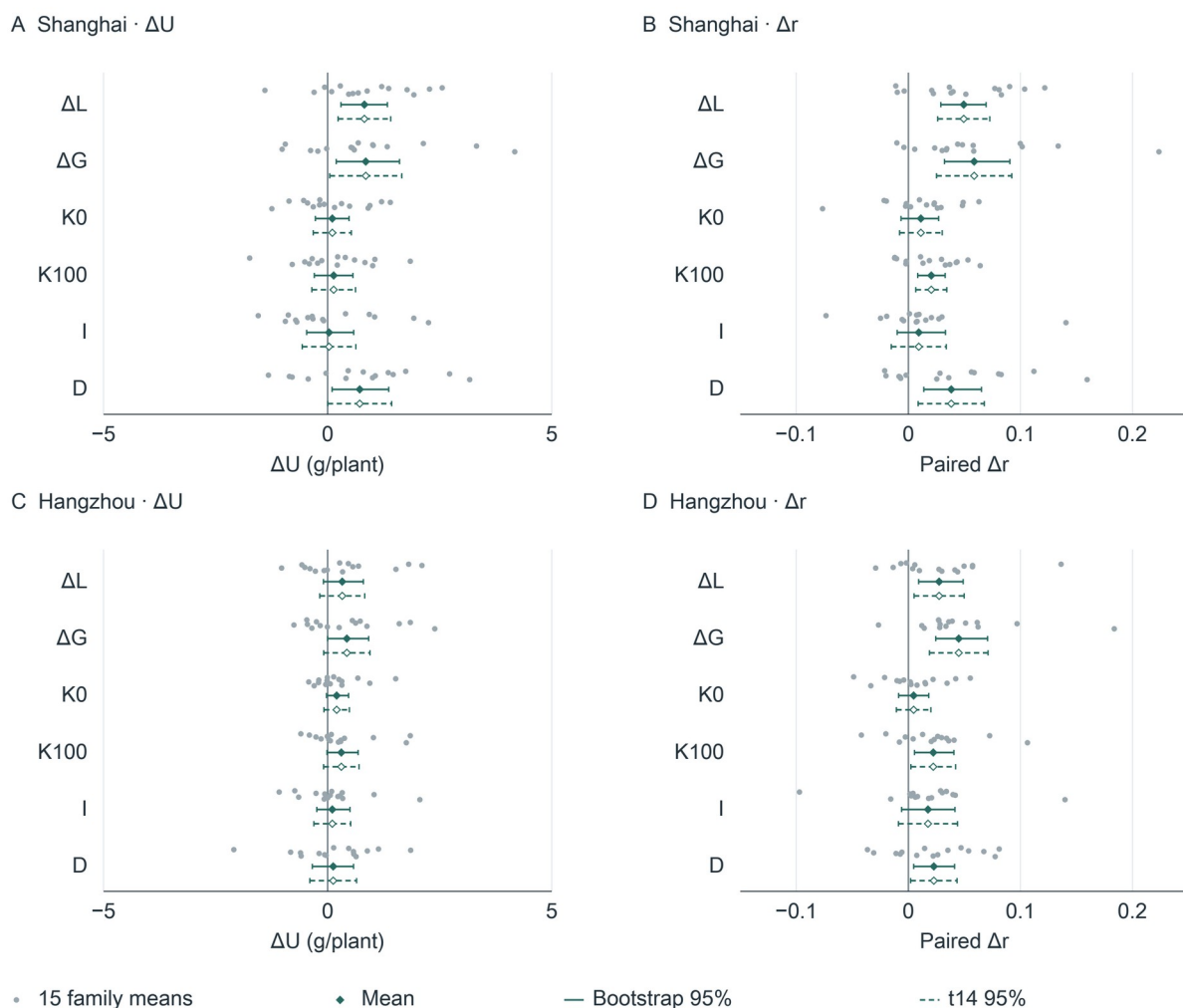
Original fixed-allocation reference: five arms $\times$ four seasons; correlated follow-up.

These 20 original cells are separate from the 20 later overlapping C/S configurations.

535
536 Figure S9. Original IRRI allocation. Five arms retain four-season U and r; only 2011 DS pool25 –
537 source Gaussian0 is primary. Later seasons are correlated.



538
 539 Figure S10. Complete MAGIC contrasts. Open/filled markers show original/no-high25 domains at
 540 SH/HZ; gray links are changes, not intervals (g/plant).



Within each contrast: family points, bootstrap mean/interval, t14 mean/interval.

Intervals are pointwise and conditional; the six contrasts share an axis within metric.

541
 542 Figure S11. NAM paired strategy contrasts. A/B: SH $\Delta U/\Delta r$; C/D: HZ $\Delta U/\Delta r$. Points show 15 five-
 543 draw family contrasts. Mean markers and solid conditional percentile intervals are separated
 544 vertically from the dashed t14 sensitivity intervals; both methods refer to the same family means.
 545 Contrasts include ΔL , ΔG , K0, K100, I and D. Intervals are pointwise; the selected-six r multiplicity
 546 sensitivity is in Section S7.