ISSN (Print): 2959-4391 ISSN (Online): 2959-4405

JOURNAL OF SOCIAL SCIENCES DEVELOPMENT

www.jssd.org.pk editor@jssd.org.pk

WEATHER FORECASTS AS A CREDENCE GOOD: VERIFIED ACCURACY, FARMER BELIEF AND THE DEMAND FOR AGROMETEOROLOGICAL INFORMATION IN PAKISTAN

Abdul Basit[1] & Nasir Iqbal[2]

[1]PhD Scholar, PIDE School of Economics, Pakistan Institute of Development Economics, Islamabad, Email: hadeedian97057@gmail.com

[2]Professor of Economics, Pakistan Institute of Development Economics, Islamabad, Email: nasir@pide.org.pk

KEYWORDS

Credence goods, information asymmetry, weather forecasts, Bayesian learning, agricultural extension, Pakistan.

ARTICLE HISTORY

Date of Submission: 26-07-2026

Date of Acceptance: 28-08-2026

Date of Publication: 03-09-2026

ABSTRACT

A farmer who acts on a weather advisory and still harvests a poor crop cannot tell a bad forecast from bad luck. Forecast quality is consumed jointly with soil, seed, pests and the weather draw, so consumption does not reveal it. Almost every study of weather-information uptake observes what farmers believe about forecast quality but not what that quality is. This paper measures both sides at the same place and the same time. Verified day-ahead accuracy is constructed for 89 Pakistan Meteorological Department stations from 34,610 scored station-days and matched to the perceived accuracy reported by 689 farmers. The level comparison depends on which verification score is used. Against percentage correct, mean belief of 0.692 sits 0.19 below a verified 0.883, but percentage correct is beaten on this panel by a no-skill climatological rule, and the sign of the gap reverses against the hit rate on rain days. What does not depend on the score is the decoupling. Across seven verification metrics the correlation between belief and verified quality never exceeds 0.112 in magnitude at station level or 0.056 at farmer level. Demand follows belief alone. Conditional on the perception, verified accuracy has no detectable effect on willingness to pay or reported outcomes. A learning model with misattribution, at calibrated coding thresholds, reproduces about two thirds of the level gap measured by percentage correct. Its policy ranking places de-biasing extension and localised advisories far above subsidised access.

Corresponding Author Abdul Basit

Email: hadeedian97057@gmail.com

DOI https://doi.org/10.59075/jssd.v5i7.259

Volume/Issue 5(7), 2026, 31-59

Introduction

The Pakistan Meteorological Department (PMD) is the national meteorological service of Pakistan. It operates a dense observatory network, produces the country's weather forecasts, and issues agrometeorological advisories through bulletins and press releases, a mobile application, a website, a helpline and the provincial agriculture extension departments. Verification of archived

day-ahead guidance at 89 of its stations puts mean accuracy at 0.879 by percentage correct, with a Heidke skill score of 0.405 that is positive at 87 of the 89. Uptake of the advisories is nevertheless low, and farmers who receive them often distrust them. A service with measurable skill is being consumed as though it had none.

This paper argues that the puzzle is an information problem with a precise name. Darby and Karni (1973) defined credence qualities as those a consumer cannot verify even after consumption. Such qualities arise, in their words, "whenever a good is utilized either in combination with other goods of uncertain properties to produce measurable output or in a production process in which output, at least in a subjective sense, is stochastic." A weather forecast is the canonical case. Its quality reaches the farmer only through the harvest, and the harvest also loads soil, seed, pest pressure, input timing and the weather draw itself. Nelson (1970) separated search from experience qualities and Stigler (1961) priced the cost of acquiring information. A credence quality is the case in which experience, however long, does not close the gap.

Akerlof (1970) showed that where sellers know quality and consumers do not, the market prices the consumer's belief. Norton (2020), whose framework this paper adapts, took that argument to fertiliser quality in Tanzanian markets. Beliefs about input quality there are wrong, persistent and decisive for willingness to pay. He explains the persistence with a learning model in which farmers misattribute stochastic crop failures to bad input quality. The setting here differs from his in two respects. There is one supplier, which gives the forecast away and has no price to signal with. And forecast quality is revealed only statistically, never in a single instance. The credence problem does not need a seller with a motive to misreport. Michelson et al. (2021) show for Tanzanian fertiliser, and Bold et al. (2017) for Ugandan inputs, that beliefs can be wrong and decisive where the input is sound. It needs only a consumer whose experience cannot identify quality.

Three further literatures bear on the design. Field evaluations of mobile and text-based agricultural advice in South Asia find small and uneven effects on practice (Fafchamps and Minten 2012; Cole and Fernando 2021). In Pakistan, Nepal et al. (2024) find no effect of weather and climate information services on farm outcomes in Punjab and Sindh, and their focus groups describe the advisories as less reliable and often unclear. The climate-services literature separates the scientific quality of a forecast from its usability and trust (Vaughan and Dessai 2014; Suckall and Bruno Soares 2020), but measures the two apart. And the learning literature shows that farmers fail to notice input dimensions their experience does not make salient (Hanna, Mullainathan and Schwartzstein 2014), which is the mechanism this paper calls misattribution. Experimental work on credence goods concentrates on the seller's incentives (Dulleck, Kerschbamer and Sutter 2011), and the disclosure literature on how quality is certified (Dranove and Jin 2010). Neither observes verified quality and belief in the same market at the same time, which is what the present design does.

The design measures both sides of the asymmetry at the local level. On the supplier side, day-ahead output of an operational global numerical weather prediction model is scored against PMD's own rain gauges, station by station, over the thirteen months before the interviews. That is

the information a meteorologist can compute and a farmer cannot. On the consumer side, perceived accuracy is elicited from farmers matched to those stations by repaired GPS coordinates. The distance between the two, farmer by farmer, is the information gap the credence property predicts.

Four findings follow. First, the level comparison is metric-dependent and is reported as such. Mean belief of 0.692 sits 0.19 below the verified accuracy of the farmers' own stations, which averages 0.883 by percentage correct. On that comparison 62.8 per cent of farmers believe less than their own station delivers. But a rule that always calls the majority class scores 0.888 on this panel, above the forecast's 0.879, so percentage correct is not by itself evidence of skill. Measured against the hit rate on rain days of 0.504, the same belief sits 0.19 above the record rather than below it. The evidence that the forecast has skill is the Heidke score, not the level of the percentage correct.

Second, and robustly, belief and verified quality are decoupled. Across seven standard verification scores the correlation between mean station belief and the station's score never exceeds 0.112 in magnitude, and at farmer level it never exceeds 0.056. Under twelve accuracy definitions the farmer-level maximum is the same 0.056. Beliefs are also more than twice as dispersed across stations as percentage correct, the score on the belief's own scale, though part of that width is the four-point scale on which beliefs are recorded. Whatever farmers are judging, it is not their own station's forecast record, however that record is scored.

Third, demand is a function of belief and of belief alone. Conditional on the perception, verified accuracy predicts neither the decision to pay, nor the amount bid, nor reported improvement in outcomes. That is the market failure the credence property implies here. A supplier who raised true accuracy would, with perceptions held fixed, see no detectable change in what farmers will pay.

Fourth, a Bayesian learning model with multiple priors and misattribution, at calibrated coding thresholds, reproduces about two thirds of the observed level gap. Its policy ranking puts de-biasing extension and village-specific advisories far above subsidised access and generic awareness campaigns. Publication of a verified accuracy record takes an intermediate place, which depends on how much weight farmers give the published number.

The setting is a useful one for the question in three respects. The supplier is a single national agency, so the quality being judged is one product rather than a mix of competing brands. The agency charges nothing, so any gap between belief and quality cannot be attributed to price signalling or to a seller's incentive to shade quality downwards. And the observatory network is dense enough, and the sampling frame close enough to it, that a farmer can be matched to the station whose forecast performance she actually experiences. The last point is what makes a local test possible. A national average of forecast skill would be compared to a national average of belief, which identifies nothing.

The contribution is a measurement one before it is a theoretical one. Studies of weather-information uptake routinely observe stated trust, and studies of forecast verification routinely observe skill, but the two are rarely observed at the same stations in the same season. Measuring them together turns the credence property from an assumption into a testable restriction. The paper also shows that the property holds where there is one honest supplier and no price. It sits on the demand side, so the remedy is not liability but disclosure, and disclosure has to be auditable to work.

Section 2 sets out the framework and its predictions. Section 3 describes the data and the construction of the two accuracy variables. Section 4 gives the empirical strategy. Section 5 reports results, Section 6 robustness, Section 7 the discussion and Section 8 concludes. Appendix A states and proves the propositions. Appendix B holds supplementary exhibits.

Conceptual framework

A credence good with a non-strategic supplier

Let the forecast system's true daily accuracy at station s be θ_s. The supplier's information set is $\mathcal{I}^S = \{\theta_s\}$. The forecaster can score every issued forecast against the gauge and knows accuracy as a number. The consumer's information set is a finite sample of remembered binary experiences together with a set of priors. The asymmetry is not that the farmer has less data. Over a season she observes many rain and no-rain outcomes. It is that her data are corrupted at source. What she codes is not whether the forecast was correct but whether her crop did well after she followed it, and that outcome also loads pests, input quality, soil and the weather draw.

Dulleck and Kerschbamer (2006) organise the credence-goods literature around a seller who knows more than the consumer and may misreport. Their conditions for the market to discipline the seller are liability, verifiability and, in extensions, reputation. None of them applies cleanly here. PMD does not sell the forecast, does not profit from over-stating it and has no diagnosis to inflate. The problem survives anyway, because it does not live in the seller's incentives. It lives in the consumer's inference.

A rain call can in principle be checked against the sky. Three things stop that check from forming the belief. The issued product is probabilistic and worded, so one dry day does not falsify a rain call. The advisory is consumed as a decision input, and it is the decision's payoff that is remembered. And if farmers did score rain calls, belief would track the hit rate or the false alarm ratio. Section 5.2 shows that it tracks neither.

Learning with multiple priors and misattribution

The farmer treats forecast quality as a Bernoulli parameter θ and holds a set of Beta priors, in the multiple-priors tradition of Gilboa and Schmeidler (1989). Each season she observes one outcome y_t, and the cohort she learns with pools such outcomes. An outcome below a threshold s_l is coded a forecast failure, one above s_h a success, and intermediate outcomes are uninformative. This is the

misattribution step, and it is what makes the good a credence good in practice. Because the outcome is noisy, the coding returns a failure even when the forecast was correct. Beliefs then update conjugately, and the set of active priors is pruned by a likelihood-ratio rule in the manner of Epstein and Schneider (2007). Ambiguity about forecast quality, as distinct from risk about the weather, is what the prior set represents. Heal and Millner (2014) make the same distinction for climate decisions. The prior set does not change the ceiling. It governs how long a range of beliefs survives before the coded data prune it, and Section 5.6 reports the single-prior case alongside.

Write $\bar{p}$ for the probability a season is coded a success and $\underline{p}$ for the probability it is coded a failure. Beliefs converge not to θ_s but to the misattribution equilibrium

$$\theta_\infty = \frac{\bar{p}}{\bar{p}+\underline{p}} \tag{1}$$

which falls short of θ_s whenever $\underline{p} > (1-\theta_s)\bar{p}/\theta_s$. Let c_t record whether the forecast acted on in season t was correct, with $\Pr(c_t = 1) = \theta_s$. Let a_c and b_c be the probabilities of a coded success and a coded failure given $c_t = c$. Then $\bar{p}(\theta_s) = a_0 + \theta_s(a_1 - a_0)$ and $\underline{p}(\theta_s) = b_0 + \theta_s(b_1 - b_0)$, and differentiating (1) gives

$$\frac{d\theta_\infty}{d\theta_s} = \frac{(a_1-a_0)\,\underline{p}-(b_1-b_0)\,\bar{p}}{\left(\bar{p}+\underline{p}\right)^2} \tag{2}$$

The numerator reduces to $a_1 b_0 - a_0 b_1$, so transmission is positive when a correct forecast raises the odds of a coded success against a coded failure. Appendix A collects the three propositions this expression supports and their proofs.

The content of the credence property is a limiting case of (2). When the coded outcome carries no information about whether the forecast was correct, so that $a_1 = a_0$ and $b_1 = b_0$, the derivative is zero. Perceived quality then does not move with true quality at all, and its level is fixed by the coding thresholds and the outcome distribution alone. Under Gaussian outcome noise, $a_1 - a_0$ and $b_1 - b_0$ both go to zero as the shift a correct forecast adds to the outcome falls relative to the noise. Transmission of true quality into belief vanishes as outcome noise grows. The good's quality is invisible to its consumer not because she has too little experience, but because experience is the wrong instrument.

Three predictions

P1, the level gap. Perceptions sit below local truth on average and consumption does not close the gap. The prediction is about a level, so it inherits the scoring rule used to define truth, and Section 5.1 reports that dependence rather than suppressing it.

P2, weak or absent transmission. In $q_i = \alpha + \beta\theta_{s(i)} + e_i$, full information implies $\beta = 1$ and a pure credence good implies $\beta = 0$. Away from the limit, misattribution learning implies a non-zero slope. It does not imply $\beta < 1$. The derivative in (2) is bounded above by $1/[4\theta_s(1-\theta_s)]$,

which exceeds one whenever θ_S differs from one half and is about 2.35 at the verified truth of 0.879. Coding environments therefore exist in which it exceeds one, so the slope on its own cannot separate misattribution from full information. The sharp prediction is instead about the unconditional association. If the coded outcome is close to uninformative about forecast correctness, belief and verified quality should be close to uncorrelated whatever the score.

P3, sufficiency. Because θ_S is not in the consumer's information set, decisions depend on θ_S only through q_i. Conditional on the perception, verified accuracy has no effect on willingness to pay or on reported value. This holds whatever the strength of transmission. The market prices the belief.

Two corollaries follow and both are tested. Among consumers facing the same θ_S, the equilibrium belief varies with the coding environment, so the cross-sectional spread of beliefs should be unrelated to the spread of verified quality. If coding environments vary more than quality does, belief will also be the more dispersed of the two. And since demand responds to θ_S only through θ_∞, an improvement in true accuracy earns the supplier little demand response. Only an intervention that changes the coding function, by raising a_1 or lowering b_1, or that shrinks outcome noise, raises the ceiling towards the truth. A change in price does neither.

Disclosure and the condition it carries

If consumption cannot reveal quality, the credence-goods remedy is disclosure rather than liability. The learning model adds a second remedy, which is to repair the coding itself, and Section 5.6 ranks the two. The supplier has to make quality observable because use of the good never will. Disclosure carries a condition, and it is the condition on which the policy argument rests. In the spirit of Dulleck and Kerschbamer (2006), and of the disclosure literature surveyed by Dranove and Jin (2010), disclosure by the supplier removes the credence problem only if the disclosed figure is itself verifiable. Here the supplier would be scoring its own product. A public agency whose budget and modernisation financing depend on demonstrated performance is not obviously indifferent to the score it publishes.

The condition is met in this case by construction rather than by assumption. The score is a percentage correct computed from two records a third party can obtain and re-derive, namely PMD's own rain-gauge register and archived model output published under an open licence. The recommendation therefore requires an auditable series rather than an announced one. The gauge record and the scored forecasts have to be released alongside the accuracy figure, at station-day resolution, so that the number can be checked instead of believed.

Data and measurement

Setting

The study population is smallholder agriculture across the four provinces of Pakistan together with Azad Jammu and Kashmir and Gilgit-Baltistan, sampled around the PMD observatory network. Farming is wheat dominated, with rice, cotton and mixed livestock. Irrigation divides the

sample in a way that matters for the argument. Canal and tube-well farmers have their drought exposure buffered by the irrigation system, so a forecast error reaches output only weakly. Rain-fed farmers carry the rainfall the forecast predicts directly into output. If experience taught quality, rain-fed farmers would be the group whose beliefs tracked verified accuracy most closely.

PMD advisories reach farmers through several channels. The department issues worded bulletins and press releases, operates a mobile application, a website and a helpline, and works through the provincial agriculture extension departments. Some respondents report receiving alerts by text message. PMD does not send those itself, so they reach farmers through third parties relaying its information. The distinction matters for the policy discussion, because it separates what the department controls from what it does not.

The two surveys

The primary sample is the AgroMet Survey, fielded around PMD observatories in June and July 2025. It was collected under the Modernization of Hydromet Services of PMD in Pakistan, a development project that fielded the instrument. Of 821 interviews one outlier is dropped, leaving 820 farmers across 110 districts. The instrument covers demographics, farm characteristics, awareness and use of PMD services, trust in forecasts, perceived value and stated willingness to pay for localised advisories. Farm size averages 5.52 acres with a median of 3.5. Owner-operators are 84.3 per cent of the sample, rain-fed farmers 32.3 per cent and wheat growers 69.0 per cent, and 61.8 per cent are aware of PMD services.

A second, independently fielded sample supports a replication. The Flood Early Warning Survey covers 850 households in 133 flood-prone districts as recorded, 117 once district names are harmonised, and concerns a different PMD product. Its median household-to-observatory distance is 6.9 km against 9.1 km in the AgroMet sample.

The sampling frame of both surveys is deliberately concentrated around PMD observatories. That concentration is what makes a station-level match possible, and it correspondingly limits how far the descriptive shares generalise to Pakistani agriculture as a whole. We acknowledge and appreciate the support of PMD and its project for provision of this data enabling this research.

What is scored, and why

PMD's data centre archives climatic data and the issued forecast through its relevant section is mainly a worded bulletin or press release. A worded forecast cannot be verified objectively in any case, as Mailier, Jolliffe and Stephenson (2006) set out. The object scored here is therefore archived day-ahead guidance from an operational global numerical weather prediction model. It is retrieved from a public historical-forecast archive under an open licence, at the coordinates of 117 PMD observatories, for 1 June 2024 to 30 June 2025. That guidance is the closest verifiable object to the issued product. It is also the closest to what the consumer encounters, since it is the kind of numerical guidance that populates the department's public forecast displays.

The window is the year before the interviews. That is the period a trust question is most likely to be answering, since it asks for a view built from recent experience.

Two objections to this choice deserve an answer. The first is that the scored object is not the issued product, so the exercise verifies a model rather than a department. The reply is that no verifiable version of the issued product exists. A categorical archive of PMD's own forecasts would be better, and creating one is a recommendation of this paper, but it does not exist for the period the beliefs refer to. The second objection is that the guidance scored here is more accurate, or less, than what a farmer actually received. The relevant claim for the argument is weaker than either. It is that the scored series orders stations in the way the issued product does, because the identifying variation in Section 5.3 is cross-station. A farmer who receives a bulletin built on guidance that is poor at her station is receiving a poor bulletin.

Constructing verified accuracy

Each observatory's coordinates are queried in the archive for the day-ahead precipitation probability and precipitation sum. Each observatory is linked to its rain gauge by a name crosswalk with manually verified overrides, since many sites carry different survey and gauge toponyms. Each station-day is then classified by the binary contrast the consumer experiences. The forecast calls rain when the probability is at least 50 per cent, the gauge records rain above 0.1 mm, and the correctness indicator equals one when the two agree. Station accuracy is the mean of that indicator,

$$\theta_s = \frac{1}{T_s} \sum_{t=1}^{T_s} \mathbb{1}\left[\text{forecast rain}_{st} = \text{actual rain}_{st}\right] \tag{3}$$

Raw scoring yields 42,033 rows. Fifteen gauges are linked to more than one observatory coordinate, so their days are scored more than once. Averaging each gauge-day to a single row leaves 34,621 unique gauge-days, and station accuracies are numerically unchanged by that collapse, with a maximum absolute change of 0.000. Restricting to stations with at least 120 verified days leaves 34,610 scored station-days at 89 stations, out of 90 gauges in the scored panel.

Mean verified accuracy is 0.879 with a standard deviation of 0.080 across the 89 stations. Farmers are matched to stations by repaired GPS coordinates. Of the 820 farmers, 689 match a scored station. They sit at 69 stations in 97 districts, and 51 stations carry at least five matched farmers. The 131 unmatched farmers all sit nearest an observatory whose gauge is not in the scored panel. They hold lower beliefs, 0.651 against 0.692, farm smaller plots, 3.5 against 5.9 acres, and are more often rain-fed, 0.40 against 0.31, so the results that follow are conditional on the matched sample.

The belief measure

Perceived accuracy q_i is the farmer's response to a four-point trust question about PMD forecasts, mapped onto the unit interval as 0, 0.25, 0.60 and 1.00. The mapping is a cardinalisation of an

ordinal scale, and Section 5.3 therefore reports an interval-censored maximum likelihood estimator alongside ordinary least squares. That estimator treats each trust category as a fixed bracket of latent accuracy, so it imposes bracket bounds rather than point values. An ordered probit with free cutpoints, which imposes no cardinal reading, is reported in Appendix B. Mean perceived accuracy is 0.685 over all 820 farmers and 0.692 over the 689 matched to a scored station.

Two features of the belief measure should be stated at the outset. The scale has four points, so part of the cross-sectional spread of q_i is an artefact of coarse recording. Comparisons of the spread of belief with the spread of verified accuracy are therefore reported as indicative rather than as tests. And q_i is a stated attitude, which is why Section 5.5 rests its decomposition on regressors measured outside the interview. Manski (2004) sets out why a coarse categorical elicitation of a subjective probability carries less information than a numerical one, and a future round of this survey should elicit beliefs as numbers with certainty follow-ups.

The variables that measure demand come from the same instrument. Willingness to pay is elicited in two parts, a participation question and an amount in rupees per season for a localised advisory. Use intensity counts how many of six listed decisions the farmer reports taking with the help of PMD information. Reported outcome improvement is a binary item on whether the respondent believes the information improved her results. The first two are behavioural in form and the last is an attitude, and the sufficiency result of Section 5.4 holds on all of them.

Cleaning

Province, district and observatory names arrive in inconsistent case, spelling and punctuation, so all joins run on a normalised key. Two records have latitude and longitude transposed, which is unambiguous because Pakistan's latitude and longitude bands barely overlap, and are swapped back. One record with an impossible longitude is excluded from distance-based matching. Left unrepaired, these records match to observatories far outside the province in which the interview took place.

A share of the tenure responses are respondent names or other free text rather than a tenure category. Pattern matching is widened to common misspellings and the remaining unmatchable entries are left missing with a flag retained as a control, because mode imputation would otherwise fabricate tenure for every respondent whose answer was unreadable. Elsewhere, categorical variables use district-stratified mode imputation and continuous variables a k-nearest-neighbour rule, with missingness indicators retained and never entered as regressors.

One further anomaly is handled openly. Every willingness-to-pay bid above a high threshold originates in a single district, whose median bid is many times the national one. That is an enumerator unit effect, and a district-stratified outlier rule cannot catch it because the whole stratum is affected. Bids are screened with a pooled log-bid rule, and every model involving bids is re-run with the flagged district excluded. Table 1 summarises the panel, the samples and the two sides of the asymmetry.

Table 1: The verification panel, the samples and the two sides of the asymmetry

Panel A. Verification panel	
Observatory coordinates scored	117
Raw scored rows	42,033
Unique gauge-days after de-duplication	34,621
Scored station-days at stations with at least 120 days	34,610
Stations	89
Verified accuracy, station mean (percentage correct)	0.879
Standard deviation across stations	0.080
Heidke skill score, station mean	0.405
Stations with a positive Heidke score	87 of 89
Panel B. Farmer samples	
AgroMet interviews analysed	820
Districts	110
Farmers matched to a scored station	689
Stations with at least one matched farmer	69
Districts in the matched sample	97
Stations with at least five matched farmers	51
Median farmer-to-observatory distance (km)	9.1
Farm size, mean (acres)	5.52
Farm size, median (acres)	3.5
Owner-operated	0.843
Rain-fed dependent	0.323
Grows wheat	0.690
Aware of PMD services	0.618
Flood Early Warning Survey households (districts)	850 (117)
Panel C. The asymmetry, 689 matched farmers	
Verified accuracy of own station, mean	0.883
Perceived accuracy, mean	0.692
Level gap against percentage correct	0.19
Share believing less than their own station delivers	0.628
Correlation of belief with verified accuracy	0.042
Standard deviation of belief across the 51 stations	0.179
Standard deviation of verified accuracy across the 51 stations	0.077
Range of belief across the 51 stations	0.639
Range of verified accuracy across the 51 stations	0.290

Empirical strategy

Three blocks follow the three predictions, and the inference plan is stated before the estimates rather than after them.

Belief formation. The first block estimates

$$q_i = \alpha + \beta\, \theta_{s(i)} + \mathbf{X}_i'\boldsymbol{\gamma} + \varepsilon_i \tag{4}$$

where $\mathbf{X}$ contains log rainfall variance at the station, education, farm size, tenure, rain-fed status, crop, livestock, log distance to the observatory and province dummies. The parameter of interest is tested against both boundaries, $\beta = 0$ and $\beta = 1$. Three estimators are reported. Ordinary least squares comes first. Because q_i derives from an ordinal scale, an interval-censored maximum likelihood estimator treats each trust category as a bracket of latent accuracy. A third column replaces numeric θ_s with a binary indicator for above-median accuracy, which asks which representation of quality the consumer processes.

Inference on β is the delicate part, because θ_s varies only across stations. Standard errors are clustered at the district level in the tables. Three design-based or robust alternatives are reported alongside. A null-imposed wild cluster bootstrap with Rademacher weights and 999 draws follows Cameron, Gelbach and Miller (2008), clustered on the 69 stations because that is where the regressor varies. Conley (1999) spatial standard errors with a Bartlett kernel and a 100 km cutoff allow for spatial dependence that district clusters do not capture. And a permutation test shuffles station accuracies, both across all stations and within province. The within-province version is the appropriate one, because (4) absorbs province dummies and the slope is identified within province. A free shuffle makes θ_s orthogonal to province in every draw and removes the collinearity that makes the real standard error large. Both are reported.

A generated-regressor problem also has to be faced. Verified accuracy is estimated from a finite number of station-days, so sampling error in it attenuates the slope towards zero. Attenuation is inconvenient here in a particular way, because zero is one of the two hypotheses of interest, and a measurement problem that pushes the estimate towards the credence boundary would flatter the theory rather than challenge it. The de-duplicated panel gives a large number of verified days at every station kept, so the sampling variance of θ_s is small relative to its variation across stations. Section 5.4 reports the reliability after the controls are partialled out and a split-sample instrumental-variable estimate that corrects for it.

Sufficiency. The second block regresses each decision outcome on both accuracies,

$$D_i = f\left(\delta_q\, q_i + \delta_\theta\, \theta_{s(i)} + \mathbf{X}_i'\boldsymbol{\lambda}\right) + \eta_i \tag{5}$$

Under P3, $\delta_\theta = 0$ conditional on q_i. Under full information q_i is redundant instead. The control vector in this block adds two further stated-belief items, an indicator that the farmer is unsure of forecast quality and the misattribution marker, since both are components of the perception the credence property conditions on. Willingness to pay is modelled with a probit on the participation margin and a Tobit on the censored amount, following Tobin (1958) and Cragg (1971). Reported outcome improvement uses a probit and use intensity ordinary least squares. All models cluster by district. The two willingness-to-pay models exclude the nine bids from the flagged district and the 16 Gilgit-Baltistan farmers, none of whom is willing to pay, so that no province dummy predicts the outcome perfectly. That leaves 664 farmers in the valuation models and 689 in the other two.

Decomposing the gap. The third block decomposes the individual gap $\theta_{s(i)} - q_i$ using the composed-error machinery of Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977), with the frontier pinned by theory at $q^* = \theta$, and with covariates entering the mean of the one-sided term as in Battese and Coelli (1995). Farmer-level gaps are recovered by the conditional expectation of Jondrow et al. (1982). Section 5.5 explains why this block is reported as a decomposition of the gap and not as an independent test of asymmetry. The covariates are the mechanism variables of the framework, namely local rainfall variance, rain-fed exposure, a stated misattribution marker, education and log distance to the station. Rain-fed exposure and rainfall variance are measured outside the interview, the first from the irrigation question and the second from the gauge record, and the argument rests on those two.

Results

The level gap, and what it depends on

Panel C of Table 1 puts the two sides face to face for the 689 matched farmers. The verified accuracy of the farmers' own stations averages 0.883 and their perceived accuracy 0.692, a level gap of 0.19, and 62.8 per cent believe less than their own station delivers. Figure 1 shows the same comparison station by station for the 51 stations with at least five matched farmers, and mean belief sits below verified accuracy at 43 of them.

That comparison is a statement about one verification score, and the score matters. Rain fell on between 10.9 and 11.2 per cent of the station-days in the panel. A forecast that always calls the majority class therefore scores 0.888 by percentage correct, which is above the 0.879 the forecast itself achieves. Percentage correct is not, on this panel, evidence of skill. The evidence of skill is the Heidke score, which is zero for a forecast no better than chance. It averages 0.405 and is positive at 87 of the 89 stations.

The sign of the level gap also reverses with the score. Measured against the hit rate on rain days, which is 0.504, the same mean belief of 0.692 sits about 0.19 above the record rather than 0.19 below it. A farmer who judged the service by how often a rain forecast is followed by a dry day would be scoring the false alarm ratio, which is 0.518. Percentage correct is used here because it is the object the trust question asks about, since a correct dry-day forecast is acted on as much as a correct wet-day one. But the level result should be read with that ambiguity in view, and the paper does not rest on it.

The decoupling, which does not depend on the score

Table 2 reports seven standard verification scores for the same panel and their correlation with belief. The forecast's record is mixed in the way a day-ahead rain forecast usually is. It catches about half of rain days, its false alarm rate is low at 0.081 because most days are dry, and about half of its rain forecasts are followed by a dry gauge. The critical success index is 0.319.

What holds across all seven is the decoupling. The largest correlation between mean station belief and the station's score is 0.112 in magnitude at station level, and 0.056 at farmer level. Under twelve accuracy definitions the farmer-level maximum is the same 0.056. Beliefs are also more dispersed than percentage correct. Across the 51 stations the standard deviation of belief is 0.179 against 0.077 for verified accuracy, and the ranges are 0.639 and 0.290. Part of the extra width is

sampling error in station means built from as few as five farmers, and part is the four-point scale. Netting out the within-station variance leaves a between-station standard deviation of belief of 0.164, still more than twice the 0.077 of verified accuracy, so the comparison is indicative rather than a test. Even allowing for that, the pattern is the one the framework predicts. Beliefs vary with the farmer and not with the forecast.

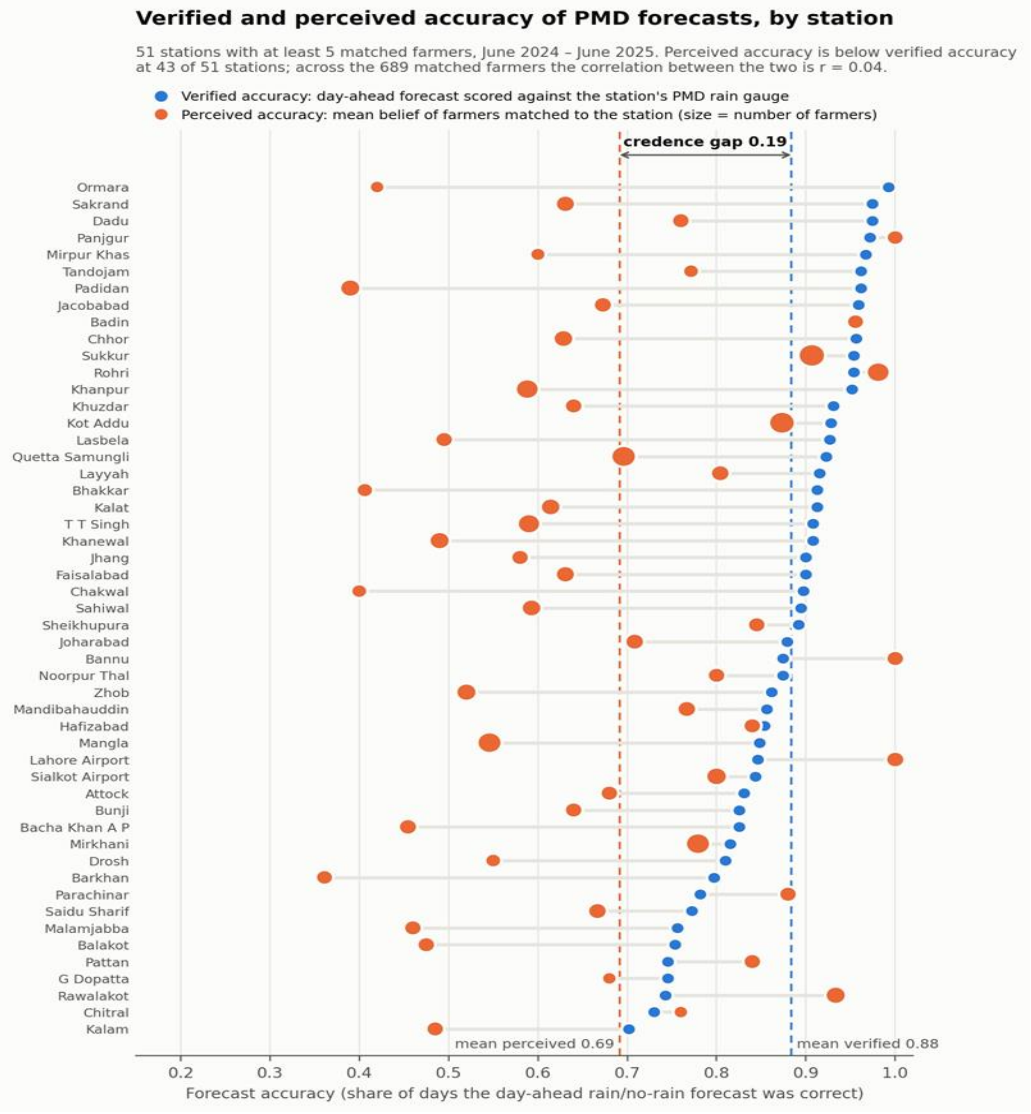

Figure 1: Verified and perceived accuracy at the 51 stations with at least five matched farmers. Verified accuracy is percentage correct, which on this panel is beaten by a climatological rule, so the level comparison shown here is metric-dependent in the way Section 5.1 describes. The robust result is the absence of any relation between the two series.

Table 2: Seven verification scores and their correlation with belief

Score	Station mean	r, 51 stations	r, 689 farmers
Percentage correct	0.879	0.051	0.042
Climatology, always the majority class	0.888	0.088	0.056
Hit rate on rain days	0.504	0.014	0.012
False alarm rate	0.081	−0.020	−0.020
False alarm ratio	0.518	0.043	0.033
Critical success index	0.319	−0.112	−0.056
Heidke skill score	0.405	−0.059	−0.023

Station means over the 89 stations with at least 120 verified days. Correlations are between the station's score and mean belief at that station (51 stations with at least five matched farmers) and between the station's score and individual belief (689 matched farmers). Hit rate is hits over hits plus misses. False alarm rate is false alarms over false alarms plus correct negatives. False alarm ratio is false alarms over hits plus false alarms. The critical success index is hits over hits plus misses plus false alarms. The Heidke score is percentage correct in excess of chance, scaled to one at a perfect forecast.

Belief formation

Table 3 estimates (4). By ordinary least squares the slope on verified accuracy is 0.983 with a district-clustered standard error of 0.470. Interval-censored maximum likelihood gives 0.905 with a standard error of 0.410. Replacing the numeric measure with a binary above-median indicator gives −0.028, which is indistinguishable from zero. The perception does not respond to a coarse classification of stations.

The slope carries less than its size suggests, and the reasons are worth setting out. Verified accuracy varies only across stations, so the effective sample is the 69 matched stations rather than 689 farmers. Clustering on the 69 stations gives $p = 0.056$, and a null-imposed wild cluster bootstrap on those clusters gives $p = 0.147$ for $\beta = 0$ and $p = 0.980$ for $\beta = 1$. Conley spatial standard errors with a 100 km kernel give 0.051 and 0.973. The within-province permutation test, which is the design-based counterpart, gives $p = 0.063$. A free permutation across all stations gives $p = 0.001$, but that shuffle makes verified accuracy orthogonal to province in every draw and so removes the collinearity the real estimate has to live with. The permutation null has a standard deviation of 0.314 under the free shuffle and 0.445 within province. The honest reading is that the conditional association is positive, cannot be distinguished from chance at the 5 per cent level, and cannot separate partial from full transmission of the ranking. Leaving out one station at a time moves the estimate over [0.708, 1.192]. Across a specification curve of estimator, clustering, controls and sample, in the manner of Simonsohn, Simmons and Nelson (2020), the median is 0.926 and every estimate is positive. Eleven of the 24 are significant at 5 per cent, and all eleven use the full sample. Removing Barkhan, the most influential station, leaves no specification significant.

Two structural covariates behave as the framework predicts and are estimated far more precisely than the slope. A log point of local rainfall variance raises belief by 0.081, and rain-fed dependence lowers it by 0.161. Wheat growers hold slightly higher beliefs. Education and log distance to the station do nothing at

all. Beliefs do not even degrade with the spatial relevance of the signal.

The province dummies are large and negative, and their pattern explains why a positive adjusted slope sits beside a raw correlation of 0.042. Across the six provinces, mean verified accuracy and mean belief run in opposite directions, with a correlation of province means of −0.465. That correlation rests on six points, its t statistic is −1.05 and its p value about 0.35, so the sign is indicative and not a finding. Balochistan has verified accuracy of 0.921 and belief of 0.625; Azad Jammu and Kashmir has 0.744 and 0.870. Within provinces the two move weakly together, with a correlation of 0.095 and a slope of 0.707. The province dummies remove the between-province part, which is why the adjusted slope is positive while the pooled correlation is flat.

Sufficiency

Table 4 estimates (5) with both accuracies in every model. The pattern is clean on three of the four outcomes. Willingness to pay responds to the perception on both the participation margin and the censored amount, while verified accuracy is conditionally null in both valuation models. Reported outcome improvement behaves the same way, with the perception at $p < 0.001$ and verified accuracy at $p = 0.392$. A supplier who improved true accuracy would, holding perceptions fixed, see no detectable change in what farmers will pay. That is a null with a wide interval rather than a proven zero, and it is the defining market failure of a credence good.

Table 3: Belief formation, perceived accuracy on verified accuracy

Regressor	(1) OLS, numeric	(2) OLS, binary	(3) Interval ML
Verified accuracy (numeric)	0.983**	—	0.905**
	(0.470)	—	(0.410)
Verified accuracy (above median)	—	−0.028	—
	—	(0.051)	—
Log rainfall variance	0.081***	0.043	0.073***
	(0.029)	(0.030)	(0.025)
Rain-fed dependent	−0.161***	−0.169***	−0.140***
	(0.032)	(0.031)	(0.028)
Grows wheat	0.070*	0.083**	0.063*
	(0.040)	(0.040)	(0.034)
Education	0.015	0.008	0.011
	(0.019)	(0.018)	(0.016)
Log distance to station	0.007	0.010	0.007
	(0.012)	(0.013)	(0.011)
Observations	689	689	689
R-squared	0.122	0.113	—

Dependent variable is perceived accuracy. District-clustered standard errors in parentheses. Province dummies, farm size, tenure and livestock included and not shown. Province coefficients are large and negative, running from −0.197 in Gilgit-Baltistan to −0.372 in Punjab, with Azad Jammu and Kashmir omitted. The interval estimator in column 3 uses the brackets 0 to 0.10, 0.10 to 0.45, 0.45 to 0.75 and

0.85 to 1.00 for the four trust categories. Four farmers whose trust response was imputed by district mode are retained; dropping them moves the slope to 0.999 and no other coefficient by more than 0.016. Stars denote significance at 10, 5 and 1 per cent under district clustering; Section 5.3 reports the wild bootstrap and permutation alternatives, under which the slope is not distinguishable from zero at 5 per cent.

The natural objection is that verified accuracy is conditionally null only because it is measured with error. Attenuation is present but too small to manufacture the null. Verified accuracy is estimated from the full panel of verified days at each station, so its sampling variance is a small fraction of its variance across stations. The reliability of verified accuracy is 0.953 in the raw cross section, 0.822 after the province dummies and 0.715 after the province dummies and farm covariates, treating station-days as independent. Attenuation of at most 28 per cent cannot turn coefficients with p above 0.39 into rejections. A split-sample estimate that instruments odd-day accuracy with even-day accuracy gives a belief-formation slope of 0.692 with a bootstrap standard error of 0.763, smaller than the least-squares slope rather than larger, so sampling error in verified accuracy is not what limits the slope. Clustering at the station level, where the regressor varies, leaves the verified-accuracy p values at 0.78, 0.68 and 0.35 in the three models with a null.

It is worth being clear about what the sufficiency result does and does not establish. It does not show that verified accuracy has no effect on demand. It shows that any such effect is not detectable once belief is controlled for, in a sample of this size, with a regressor that varies across stations rather than across farmers. That is the empirical content the credence property has in a cross section. The complementary evidence is that the same regressor also fails to predict belief itself, so the channel through which it would have to act is missing as well.

One outcome departs from the pattern and it must be reported carefully. Use intensity responds to both regressors in this specification, with a coefficient on verified accuracy of 6.314 at $p = 0.0003$. The rejection survives a Holm correction across the four outcomes, with an adjusted p of 0.0012, so it is not a multiplicity artefact and has to be confronted on specification. That coefficient does not survive a control for rain-day frequency. Verified accuracy is a station-level variable that is highest on the arid coast and lowest in the mountains, and irrigation regime, cropping calendar and extension density all shift along the same gradient. Adding rain-day frequency cuts the coefficient to 1.884 with $p = 0.538$. Adding altitude alone raises it to 7.284 with $p = 0.010$. Adding rain-day frequency and altitude together leaves 2.876 with $p = 0.431$, and district fixed effects leave 1.495 with $p = 0.799$. The perceived-quality coefficient stays between 0.94 and 0.99 with $p \le 0.0015$ throughout. Table 6 in Appendix B reports the sequence. The finding this block supports is therefore the null on verified accuracy, and no behavioural reading is placed on the 6.314.

Decomposing the gap

The individual gap can be decomposed with a composed-error model whose frontier theory pins at $q^{*} = \theta$. Estimated that way, a half-normal specification splits the gap almost evenly, with a one-sided standard deviation of 0.248 against a symmetric one of 0.234, a ratio of 1.06, and

Table 4: Sufficiency, decision outcomes on both accuracies

Outcome and model	q coef.	q p	θ coef.	θ p
Probit, willing to pay	0.958	0.001	−1.394	0.748
Tobit, stated WTP (hundreds PKR)	3.220	0.006	−5.064	0.647
OLS, use intensity (0 to 6)	0.994	0.000	6.314	0.0003
Probit, reports improved outcomes	1.726	0.000	2.769	0.392

Each row is one model containing the perception, verified accuracy, a stated ambiguity indicator, the stated misattribution marker, education, farm size, tenure, rain-fed status, log rainfall variance and province dummies, with district-clustered standard errors over 97 districts. The willingness-to-pay rows exclude the flagged district's nine bids and the 16 Gilgit-Baltistan farmers, none of whom is willing to pay, leaving 664; the other rows use all 689. With the nine bids included the perception coefficients are 0.986 (p = 0.001) and 26.29 (p = 0.104) and the verified-accuracy coefficients −1.197 (p = 0.783) and 14.91 (p = 0.883). The credence prediction is that verified accuracy is conditionally null. The use-intensity coefficient on verified accuracy does not survive agro-ecological controls, and Table 6 reports what happens to it.

An implied mean one-sided gap of 0.198. Letting the frontier intercept float upward by 0.216 produces a much larger ratio of 8.28 and a one-sided standard deviation of 0.492, but it puts the full-information benchmark above one, the highest accuracy a forecast can have, for 87 per cent of matched farmers. When the mean of the one-sided term is allowed to depend on covariates, its dispersion parameter runs to zero, so the fitted values are a linear decomposition of the gap rather than a recovered one-sided error. The standard errors on the covariates below are from a district-clustered bootstrap of that likelihood with 200 draws.

How much of the gap the model calls one-sided therefore depends on where the frontier is placed. This block is reported as a decomposition of the gap and not as a third independent test of asymmetry. Two further facts point the same way. The one-sided restriction is imposed against the framework's own corollary, which predicts dispersion in both directions, and against the data, in which 37.2 per cent of matched farmers sit above their own station.

What the decomposition adds is the covariate pattern. A stated misattribution marker adds 0.243 to the gap with a standard error of 0.032, rain-fed exposure adds 0.136 with a standard error of 0.039, and a log point of rainfall variance reduces it by 0.080 with a standard error of 0.023. Education at −0.021 and log distance at 0.016 are not distinguishable from zero.

The misattribution coefficient needs a qualification the other two do not. The dependent variable is the gap between verified accuracy and the farmer's stated view of quality, and the misattribution indicator is a second stated view of quality collected from the same respondent in the same interview. Regressor and outcome share a source. The cell is also thin, since misattribution equals one for 69 of the 689 matched farmers, so the 0.243 is a conditional association among about seventy people. Rain-fed status comes from the irrigation question and rainfall variance from the gauge record. Those two are

measured outside the interview and carry the part of the decomposition that is free of the common-source objection.

The learning model and the policy simulations

The learning model of Section 2 is simulated with the verified truth of 0.879 as the target. A cohort of 52 farmers pools its coded outcomes over 25 seasons. A season in the worst 15 per cent of outcomes is coded a forecast failure and one at or above the median a success, following the base paper's calibration of an oversold input. The cohort holds nine Beta priors with means from 0.1 to 0.9, prunes them at a likelihood ratio of 0.10, and each scenario is replicated 1,000 times. In the baseline the coded outcome carries no information about forecast correctness, so the model sits at the limit of Proposition 1 and the analytic ceiling is 0.50 divided by 0.65, or 0.769. Beliefs settle near that ceiling, at a simulated mean of 0.757 and a mean distance from the truth of 0.122, against an observed matched-sample mean of 0.692. Against the station-mean truth the observed gap is 0.187, so the ceiling reproduces about two thirds of it, or 65 per cent, and leaves the rest unexplained. That share is set by where the coding thresholds are placed. A failure threshold at the 10th percentile gives a ceiling of 0.833 and one at the 20th percentile 0.714, so the two thirds is a calibration output rather than a test. The single-prior case settles at 0.768, about one point above the baseline, which is what the prior set adds to the level. What the remaining third is cannot be settled from this design. The four-point trust scale, the response style of the interview and the wording of the elicitation are all candidates and none is established here. The model is not claimed to reproduce the belief distribution, and the two-thirds figure is not a test of the model, since the thresholds that produce it are calibrated rather than estimated.

Table 5 reports the scenarios as the share of the distance to the truth each one closes, and the ranking has to be read with the way each scenario is coded in view. Free alerts close 2.3 per cent and an information campaign 7.2 per cent. Both are low for the same structural reason. The model contains no channel through which wider coverage or generic awareness changes anyone's inference. Village-specific advisories close 50.8 per cent. The scenario is coded as a fall in the share of seasons read as forecast failures, from the worst 15 per cent of outcomes to the worst 10 per cent, which is what a more local advisory does at a fixed failure threshold when it reduces the noise in which the signal is embedded.

De-biasing is the other lever, and the figure to plan against is the partial one. Removing misattribution entirely closes 90.8 per cent of the distance, but misattribution is the model's only source of belief bias, so a scenario that switches it off must remove almost the whole modelled gap. That number is an upper bound by construction and is reported as one. A programme that reaches half of the farmers in a cohort, so that half of the pooled coded seasons are read correctly, closes 56.2 per cent, which is close to what localisation achieves and is the defensible planning figure.

Publication of a verified accuracy record is a different instrument and needs its own scenario. It does not change how a farmer codes a season. It supplies a signal about verified accuracy from outside the coding channel, so it enters the belief update alongside the coded seasons. Its strength is the number of informative trials a season the published record is worth, against about 34 coded trials the cohort generates

each season. At one trial it closes 3.2 per cent, at five 12.4 per cent and at twenty 37.1 per cent. Publication beats subsidised access at every weight tried, and beats a generic campaign once the record is worth about five trials a season. At the weights tried it does not reach the localisation or de-biasing scenarios, but a published record covering a year of station-days is worth far more than twenty trials, so the last row understates what a credible record could do. The model has no theory of the weight a farmer gives a published number, so the ranking of this instrument is the least disciplined in the table. Its case therefore rests on the sufficiency result of Section 5.4 and the decoupling of Section 5.2 as much as on this ranking, since a quality the demand side cannot see is a quality no amount of access will price.

Table 5: Policy simulations from the calibrated learning model

Scenario	Distance closed (%)
Free alerts or subsidy, 50 per cent more users	2.3
Information campaign	7.2
Published accuracy record, worth 1 trial a season	3.2
Published accuracy record, worth 5 trials a season	12.4
Published accuracy record, worth 20 trials a season	37.1
Village-specific advisories, failure threshold 15th to 10th pct.	50.8
Partial de-biasing, half the misattribution removed	56.2
Full de-biasing, no misattribution	90.8

Distance is measured from a verified truth of 0.879. The baseline is a mean terminal belief of 0.757 and a distance of 0.122. The full de-biasing row switches off the model's only source of belief bias and is an upper bound by construction. The three publication rows add w informative trials at the true accuracy to each season's coded outcomes, against about 34 coded trials a season, and are the only scenarios representing published verification.

Robustness

The cardinalisation of the trust scale. Perceived accuracy is a four-point ordinal response mapped onto the unit interval, and every level statement in Section 5.1 depends on that mapping. Column 3 of Table 3 re-estimates the belief-formation equation with an interval-censored maximum likelihood estimator that replaces point values with brackets, treating each trust category as a bracket of a latent accuracy. It reproduces the slope and the pattern of the covariates. An ordered probit of the raw trust category on verified accuracy, which imposes no cardinal reading, gives a coefficient of 4.234 with a station-bootstrap p of 0.141, the same picture in a different metric. The decoupling is robust to the mapping for a simpler reason. Any monotone remapping leaves the rank ordering of beliefs unchanged, and the association between belief and verified quality is near zero at every station and every farmer under every metric tried, so there is no association for a remapping to strengthen. The level gap is the one result that carries the cardinalisation, which is a second reason, alongside the base-rate problem, for not resting the argument on it.

Alternative definitions of quality. Section 5.2 reported seven verification scores. Twelve accuracy

definitions in total have been tried. They are the seven scores of Table 2, an above-median indicator, the hit rate when the forecast total itself exceeds 1.0 mm, its above-median indicator, and the mean absolute and root mean squared error of the forecast total in millimetres. The farmer-level correlation between belief and objective quality never exceeds 0.056 in magnitude under any of them. The decoupling is not an artefact of one threshold choice. The adjusted slope of Table 3 is less robust. Under the 1.0 mm definition it falls to 0.087 with $p = 0.87$, under mean absolute error it is 0.077 per standard deviation with $p = 0.03$, and under root mean squared error 0.009 with $p = 0.78$. The slope is reported as suggestive for that reason as well.

Design-based inference and falsification. Table 7 in Appendix B collects the diagnostics on the belief-formation slope, which is the paper's most delicately identified parameter. The within-province permutation gives $p = 0.063$ and the wild cluster bootstrap $p = 0.147$, so the design-based and model-based families agree. A placebo using the second-nearest scored station gives a coefficient of −0.116 with $p = 0.870$, and in a horse race the own-station coefficient is 1.235 with $p = 0.037$ against −0.606 with $p = 0.449$ for the neighbour. Whatever the slope is picking up, it is not correlated regional sentiment.

Base-rate dependence. The base rate of rain on this panel is between 0.109 and 0.112, so percentage correct is high wherever rain is rare, as Mailier, Jolliffe and Stephenson (2006) warn. That is why Section 5.1 reports the climatological benchmark of 0.888 and the reversal against the hit rate, and why the argument rests on the decoupling and on sufficiency rather than on the level.

Monthly wedges. Pooling the de-duplicated station-days by month gives thirteen monthly accuracy figures. Mean belief over all 820 farmers of 0.685, since no station match is needed for a comparison with a panel mean, with a district-clustered standard error of 0.021 over 110 districts, sits below twelve of the thirteen and outside the band in twelve. The maximum clustered t statistic is 12.9. The exception is August 2024, the worst monsoon month, where verified accuracy is 0.680 and the wedge is −0.005 with $t = -0.19$. These are not thirteen pieces of evidence. Every comparison uses the same mean belief, and the monthly accuracy figures are built from overlapping stations, so this is one comparison repeated thirteen times.

Enumerator effects and one verification year. All willingness-to-pay estimates exclude the flagged district's bids, and within-district estimates of the belief-to-bid gradient reproduce the positive belief-to-bid gradient. A separate concern is temporal. Verified accuracy is built from a single window, while belief is the residue of a longer history. If that year were unusually good or bad for Pakistani forecasting, the level comparison would not be like for like. Splitting the window at mid-December 2024 gives station means of 0.856 and 0.901 in the two halves, with a Spearman correlation of 0.88 across the 89 stations, so the ranking of stations is stable within the year even though the level moves. Two things limit the damage without removing it. The identifying variation in Section 5.3 is the relative skill of stations rather than the level, and the level result is already reported as the weakest of the three. A defensible test needs the same verification exercise run over several monsoon cycles, which is a data-acquisition project rather than an extension of this one.

The flood sample. The Flood Early Warning Survey applies the same framework to a different PMD

product and an independently fielded sample of 850 households. Mean perceived accuracy is 0.71, within three points of the AgroMet figure. Households who are unsure whether they experienced a false alarm report perceived accuracy 0.091 lower than those who are sure they did not, with a standard error of 0.024. Flood frequency raises belief by 0.064 and mega-flood experience lowers it by 0.080. The reading that makes the first coefficient interesting is that less information should move beliefs towards a prior rather than below both better-informed groups. But belief and the false-alarm recall are both self-reported by one respondent in one interview, so the estimate is common-source. A household disengaged from the warning system will both recall the system's performance less well and rate it poorly, and this design cannot separate that reading from the first. The flood result is therefore corroboration from a separate sample and a different product. It is not the study's decisive estimate.

Discussion and policy

Three findings organise what follows, and they are not equally strong. The decoupling is the robust one. Belief and verified quality are unrelated under every scoring rule tried, at both levels of aggregation, and beliefs are more dispersed than quality. Sufficiency is the second. Conditional on the perception, verified accuracy predicts neither payment nor reported value. The level gap is the third and the weakest, because it inherits a scoring rule that a no-skill climatological benchmark beats on this panel, and it reverses sign against the hit rate.

Taken together they say that the level of trust is set by the coding environment and not by the record. Whether the ranking of stations transmits into belief the data cannot decide. The pooled slope of belief on verified accuracy with no controls is 0.165 with $p = 0.64$, and the covariate-adjusted slope cannot be distinguished from one. A better forecast would earn PMD no detectable rise in trust in this sample. That is an uncomfortable conclusion for a service that has invested in skill, and it is the reason the policy argument does not begin with the forecast.

For PMD and its relevant Ministry the implications are concrete. Investments that raise verified accuracy, through better models and denser observation, are worth making on agronomic grounds. On their own they will not raise adoption or willingness to pay by a detectable amount, because the demand side prices the belief. Subsidies and awareness campaigns barely move beliefs in the simulations, because they leave the faulty inference intact.

What repairs the inference is what the model rewards. Extension programmes that teach what a probabilistic forecast promises, and what a bad season does and does not imply, come first. A programme that de-biases half of the coded seasons a cohort pools closes 56.2 per cent of the distance to the truth in the model, and that is the figure to plan against. The 90.8 per cent for complete de-biasing removes the model's only bias mechanism and is an upper bound by construction. Village-specific advisories close 50.8 per cent by reducing the share of seasons that are charged to the forecast.

Publishing a simple, local, verifiable accuracy record is the second instrument, and it converts a credence attribute into a search attribute, which is the direction Darby and Karni (1973) point to. Its simulated ranking is intermediate and depends on the weight farmers give the published number, closing 3.2 per cent at one informative trial a season and 37.1 per cent at twenty, which is under two thirds of one

season's coded experience. Its stronger support is the invisibility result itself. Field evidence points the same way. Roudier et al. (2014) find that participatory work in which forecasts are explained alongside their record changes how smallholders use them, and Burgeno and Joslyn (2020) find that experienced inaccuracy erodes user trust more than inconsistency between successive forecasts does. A service whose accuracy is high but never displayed forgoes the trust that accuracy would otherwise earn. The published score should be a skill score against climatology, such as the Heidke score, rather than the share of correct days alone, for the reason Section 5.1 gives. And it has to be auditable rather than announced, which means releasing the gauge record and the scored forecasts at station–day resolution.

The third instrument is to aim the first two at the farmers who carry the largest gaps. Rain-fed producers are the group the decomposition identifies on externally measured characteristics.

There is also a warning. Overselling advisories raises the outcome farmers expect and so raises the number of seasons they code as failures, which mechanically increases misattribution. Norton (2020) draws the same caution for fertiliser certification. Honest communication of what a forecast can and cannot do is not public relations. In this model it is accuracy policy.

Two limitations bound all of this. The verification window is a single year, and beliefs are the residue of a longer history, so a multi-year verification panel is the natural extension. And the sampling frame is concentrated around PMD observatories, which is what makes the station match possible and what limits how far the descriptive shares travel. The data are a single cross section, so the decomposition reports conditional associations and the causal reading comes from the calibrated model rather than from a quasi-experiment. Beliefs were elicited on a four-point scale rather than as probabilities, and a future round should elicit them as numbers.

Conclusion

This paper tested whether weather forecasts trade as a credence good in Pakistani agricultural markets, by measuring what earlier work could only assume. The supplier's information, verified station-level accuracy, averages 0.879 by percentage correct and carries a Heidke skill score of 0.405 that is positive at 87 of 89 stations. The consumer's information, built from yes-or-no experience, averages 0.692, varies more than twice as widely across stations as percentage correct, and correlates with verified quality at 0.042. That correlation stays below 0.112 in magnitude under seven scoring rules and below 0.056 at farmer level under twelve. Decisions price the belief and take no notice of the truth once the belief is allowed for.

The level comparison depends on which score defines truth, and the paper says so. The decoupling and the sufficiency result do not. A learning model with misattribution, at calibrated coding thresholds, reproduces about two thirds of the level gap and identifies the coding of experience, rather than price or coverage, as the binding constraint on demand.

The wider point is about the credence-goods literature. In that literature the problem is a seller's incentive to misreport, and the remedies are liability, verifiability and reputation. Here the seller is honest, the good is free and the good has measurable skill, and the problem is still present, because it sits

on the demand side. Consumption cannot reveal quality, so demand cannot reward it. Ma, Spielman, Nazli, Zambrano, Zaidi and Kouser (2014) find the same structure in Pakistani Bt cotton seed, where prices bear little relation to assayed trait efficacy, and Michelson et al. (2021) and Bold et al. (2017) find it for fertiliser in East Africa. A forecast is a harder case than any of those, because its quality is revealed only statistically, across many realisations, and never in a single instance. If misattribution is this durable where the supplier is honest and the good is free, the difficulty found here is a lower bound on the difficulty of the case where the seller is adversarial. Pakistan's agrometeorological services do not have a forecasting problem. They have a credence problem, and the policy that fixes it is the one that repairs the farmer's inference.

Data availability

The survey data and daily gauge record are the property of Pakistan Meteorological Department. The archived day-ahead forecasts are the historical-forecast archive of Open-Meteo (2025), which redistributes output of the ICON model of Deutscher Wetterdienst under a CC-BY 4.0 licence.

Appendix A. Propositions and proofs

Let $c_t \in \{0,1\}$ record whether the forecast acted on in season t was correct, with $\Pr(c_t = 1) = \theta_s$. The outcome observed is drawn from a distribution depending on c_t, with a correct forecast shifting it up by $\Delta > 0$ against noise of spread σ. The farmer codes a success when the outcome exceeds s_h and a failure when it falls below s_l. Write a_c for the probability of a coded success and b_c for the probability of a coded failure given $c_t = c$. Then $\bar{p}(\theta_s) = a_0 + \theta_s(a_1 - a_0)$ and $\underline{p}(\theta_s) = b_0 + \theta_s(b_1 - b_0)$, and the equilibrium belief is $\theta_\infty = \bar{p}/\left(\bar{p} + \underline{p}\right)$ as in (1).

Proposition 1, quality invisibility. Differentiating θ_∞ with respect to θ_s gives (2). When $a_1 = a_0$ and $b_1 = b_0$, the numerator vanishes, so $d\theta_\infty/d\theta_s = 0$ and $\theta_\infty = a_0/(a_0 + b_0)$, which is fixed by the coding thresholds and the outcome distribution alone.

Proof. The derivative follows from the quotient rule applied to $\bar{p}/\left(\bar{p} + \underline{p}\right)$ with $d\bar{p}/d\theta_s = a_1 - a_0$ and $d\underline{p}/d\theta_s = b_1 - b_0$. Setting both differences to zero makes $\bar{p}$ and $\underline{p}$ constant in θ_s, which gives the level. Under Gaussian noise, $a_1 - a_0 = \Phi\big((\mu_1 - s_h)/\sigma\big) - \Phi\big((\mu_0 - s_h)/\sigma\big)$, which tends to zero as Δ/σ tends to zero, and the same holds for $b_1 - b_0$. The denominator tends to $a_0 + b_0$, the share of seasons that are coded at all, so the derivative tends to zero. Transmission of true quality into belief therefore vanishes as outcome noise grows relative to what a correct forecast adds to the outcome. □

Proposition 1 is a limiting case, and away from the limit transmission is partial. The derivative in (2) is bounded above by $1/[4\theta_s(1 - \theta_s)]$ and exceeds one in some coding environments. The bound is attained at $a_0 = b_1 = 0$, an environment with no misattribution, so a slope above one is not a signature of misattribution. The slope alone cannot separate misattribution from full information, which is why the empirical argument rests on the unconditional decoupling and on sufficiency.

Proposition 2, dispersion set by the environment. Among consumers facing the same θ_s, θ_∞ varies

with a_0 and b_0. Under the conditions of Proposition 1 the cross-sectional variance of beliefs is the variance of coding environments and is unrelated to the variance of true quality.

Proof. With $a_1 = a_0$ and $b_1 = b_0$, $\theta_\infty = a_0/(a_0 + b_0)$ does not depend on θ_s, so all cross-sectional variation in θ_∞ comes from variation in (a_0, b_0), which are properties of the consumer's thresholds and outcome distribution. The empirical signature is a belief distribution uncorrelated with verified accuracy under any verification metric. Where coding environments vary more than quality does, beliefs are also the more dispersed. □

Proposition 3, invisibility to demand when transmission is weak. Let demand be $D(q)$, increasing in perceived quality. Then $dD/d\theta_s = D'(q)\, d\theta_\infty/d\theta_s$, which is zero in the limit of Proposition 1 and small whenever transmission is weak.

Proof. Immediate from the chain rule and Proposition 1. Two consequences follow. An improvement in true accuracy earns the supplier little or no demand response. And only an intervention that raises a_1, lowers b_1, or shrinks σ raises the ceiling towards the truth. The ceiling in (1) is increasing in a_1 and decreasing in b_1. Lowering b_1, the probability that a correct forecast is charged with a failure, also raises the numerator of (2) and lowers its denominator, so transmission rises too. Teaching attribution does this by changing how outcomes are coded, and a plot-specific advisory does it by lowering b_0, the probability that a season is coded a failure. A change in price does neither and leaves θ_∞ where it was. □

Published verification works differently and is not licensed by Proposition 3. It does not change the coding function. It supplies an exogenous signal about θ_s from outside the coding channel, and the simulation in Section 5.6 models it that way, as informative trials entering the belief update alongside the coded seasons.

Appendix B. Supplementary exhibits

Table 6: Use intensity with agro-ecological controls

Specification	θ coefficient	p
As Table 4, province dummies	6.314	0.000 3
Add rain-day frequency	1.884	0.538
Add altitude	7.284	0.010
Add rain-day frequency and altitude	2.876	0.431
Add both, with district fixed effects	1.495	0.799
Published controls, with district fixed effects	−2.387	0.643

Ordinary least squares on the 689 matched farmers with district-clustered standard errors, with perceived quality, the full control set and log rainfall variance in every row. The coefficient on perceived quality stays between 0.94 and 0.99 with $p \le 0.0015$ in all four rows. The district-fixed-effects row is a weak check, because verified accuracy does not vary within a station and few districts

hold more than one scored station.

Table 7: Diagnostics on the belief-formation slope

Diagnostic	Result
District-clustered p, $\beta = 0$	0.036
Station-clustered p, $\beta = 0$ ($\beta = 1$)	0.056 (0.974)
Wild cluster bootstrap p, $\beta = 0$	0.147
Wild cluster bootstrap p, $\beta = 1$	0.980
Conley spatial HAC (100 km), p for $\beta = 0$	0.051
Conley spatial HAC (100 km), p for $\beta = 1$	0.973
Permutation p, shuffled across all stations	0.001
Permutation p, shuffled within province	0.063
Split-sample errors-in-variables 2SLS, slope (bootstrap se)	0.692 (0.763)
Ordered probit of the trust category, coefficient (station-bootstrap p)	4.234 (0.141)
Placebo, second-nearest station coefficient (p)	−0.116 (0.870)
Horse race, own station (p)	1.235 (0.037)
Horse race, second-nearest station (p)	−0.606 (0.449)
Jackknife range, leave one station out	[0.708, 1.192]
Specification curve, median	0.926
Specification curve, share positive	100 per cent
Specification curve, share significant at 5 per cent	46 per cent

The within-province permutation is the appropriate design-based test, because the slope is identified within province once province dummies are absorbed. The free permutation makes verified accuracy orthogonal to province in every draw. The wild bootstrap uses 999 Rademacher draws on the 69 station clusters. The share significant in the specification curve uses analytic district- or station-clustered p values, not the bootstrap or permutation tests, and the eleven significant specifications all use the full sample.

Table 8: Flood Early Warning Survey, perceived accuracy

Regressor	Coefficient (Standard Error)
Sure a false alarm was experienced	0.050 (0.041)
Not sure whether a false alarm was experienced	−0.091*** (0.024)
Flood frequency	0.064*** (0.016)
Mega-flood experience	−0.080** (0.034)
Observations	850
Mean perceived accuracy	0.71

Ordinary least squares with district-clustered standard errors, province dummies and controls for log rainfall variance, warning lead time and distance to the observatory, on 850 households in 117

districts. The omitted category for the first two rows is households sure they did not experience a false alarm. Belief and the false-alarm recall are both self-reported in one interview, so these estimates are common-source and corroborative only.

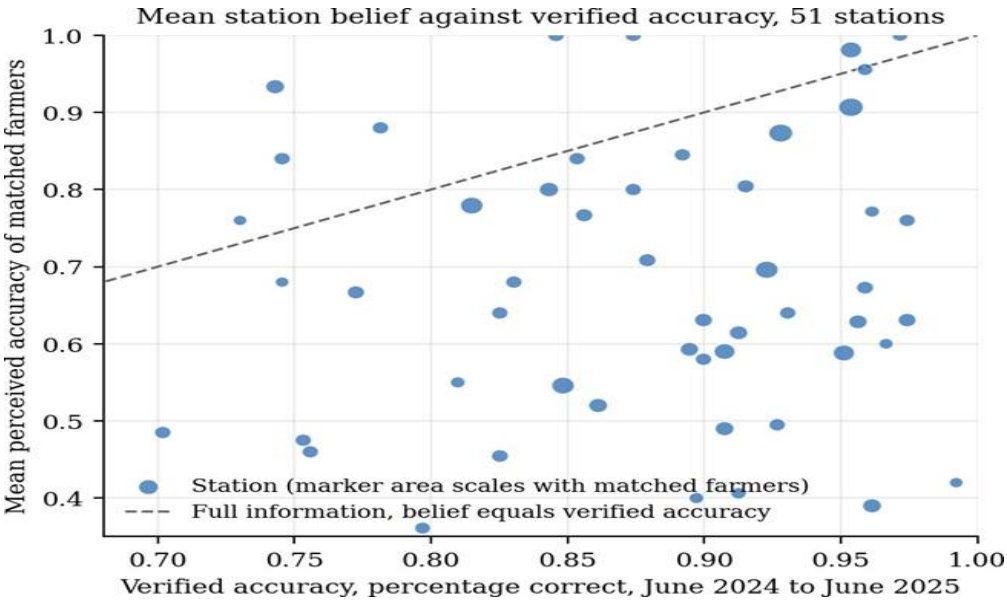

Figure 2: Mean station belief against verified accuracy at the 51 stations with at least five matched farmers, with the full-information line. Marker area scales with the number of matched farmers. The paper's estimate of the slope is the covariate-adjusted one in Table 3, and the paper does not claim that the ranking of stations transmits, because the within-province permutation does not reject at 5 per cent.

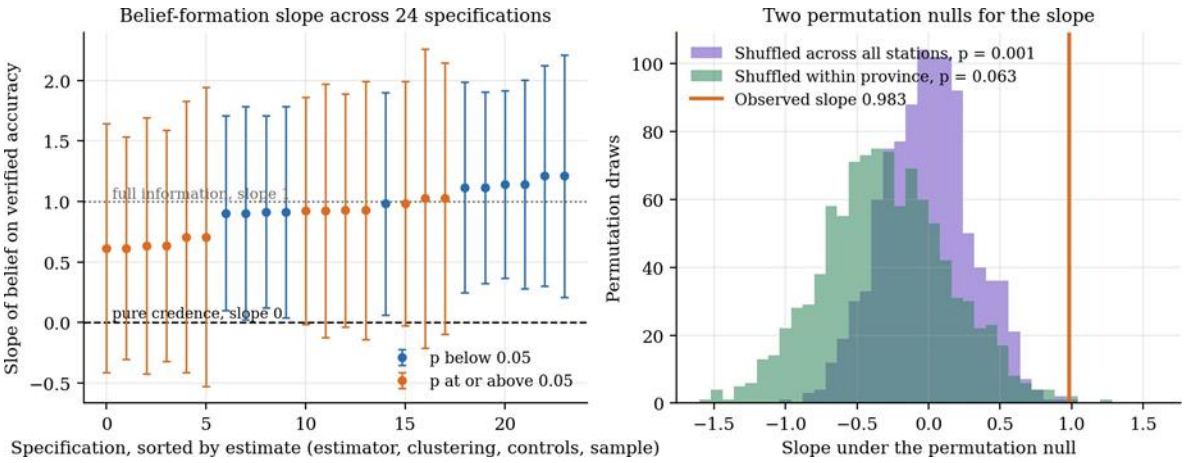

Figure 3: Specification curve and permutation tests for the belief-formation slope. The left panel shows the 24 estimates with 95 per cent intervals, sorted by size. The right panel overlays the two

permutation nulls. The free shuffle across all stations gives p = 0.001. The appropriate test shuffles within province, holding fixed the structure the province dummies absorb, and gives p = 0.063, as Table 7 reports.

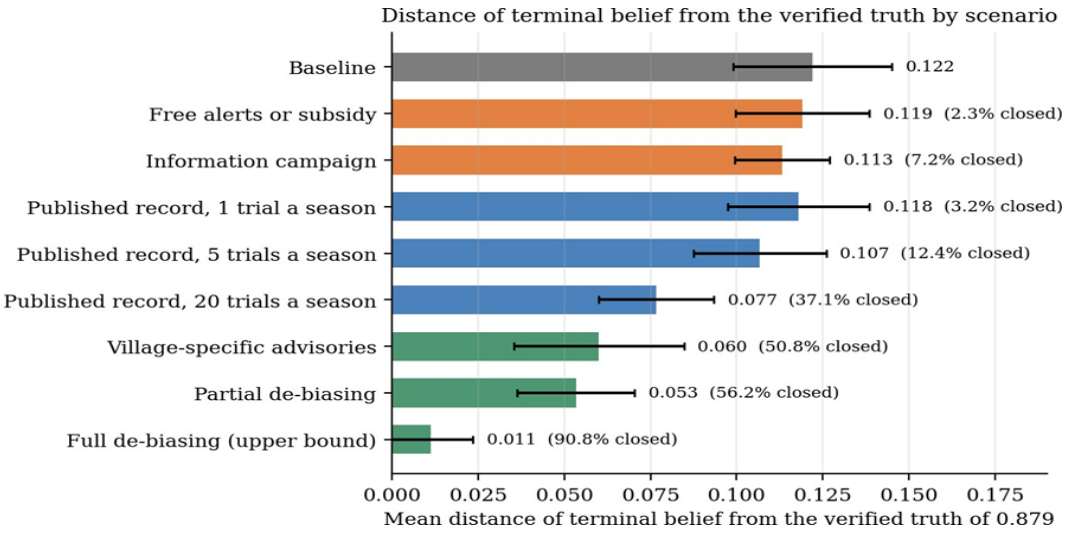

Figure 4: Simulated distance of terminal belief from the verified truth of 0.879 under the baseline and the eight scenarios of Table 5. Bars are means over 1,000 replications and whiskers one standard deviation across replications. Full de-biasing switches off the model's only source of belief bias and is an upper bound by construction; partial de-biasing, which closes 56.2 per cent, is the planning figure.

References

Aigner, D., Lovell, C.A.K. and Schmidt, P. (1977). Formulation and Estimation of Stochastic Frontier Production Function Models. Journal of Econometrics 6(1), 21–37.

Akerlof, G.A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. Quarterly Journal of Economics 84(3), 488–500.

Battese, G.E. and Coelli, T.J. (1995). A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data. Empirical Economics 20(2), 325–332.

Bold, T., Kaizzi, K.C., Svensson, J. and Yanagizawa-Drott, D. (2017). Lemon Technologies and Adoption: Measurement, Theory and Evidence from Agricultural Markets in Uganda. Quarterly Journal of Economics 132(3), 1055–1100.

Burgeno, J.N. and Joslyn, S.L. (2020). The Impact of Weather Forecast Inconsistency on User Trust. Weather, Climate, and Society 12(4), 679–693.

Cole, S. and Fernando, A.N. (2021). 'Mobile'izing Agricultural Advice: Technology Adoption, Diffusion, and Sustainability. Economic Journal 131(633), 192–219.

Cameron, A.C., Gelbach, J.B. and Miller, D.L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. Review of Economics and Statistics 90(3), 414–427.

Conley, T.G. (1999). GMM Estimation with Cross Sectional Dependence. Journal of Econometrics 92(1), 1–45.

Cragg, J.G. (1971). Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. Econometrica 39(5), 829–844.

Darby, M.R. and Karni, E. (1973). Free Competition and the Optimal Amount of Fraud. Journal of Law and Economics 16(1), 67–88.

Dranove, D. and Jin, G.Z. (2010). Quality Disclosure and Certification: Theory and Practice.

Journal of Economic Literature 48(4), 935–963.

Dulleck, U. and Kerschbamer, R. (2006). On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods. Journal of Economic Literature 44(1), 5–42.

Dulleck, U., Kerschbamer, R. and Sutter, M. (2011). The Economics of Credence Goods: An Experiment on the Role of Liability, Verifiability, Reputation, and Competition. American Economic Review 101(2), 526–555.

Epstein, L.G. and Schneider, M. (2007). Learning Under Ambiguity. Review of Economic Studies 74(4), 1275–1303.

Fafchamps, M. and Minten, B. (2012). Impact of SMS-Based Agricultural Information on Indian Farmers. World Bank Economic Review 26(3), 383–414.

Gilboa, I. and Schmeidler, D. (1989). Maxmin Expected Utility with Non-unique Prior. Journal of Mathematical Economics 18(2), 141–153.

Hanna, R., Mullainathan, S. and Schwartzstein, J. (2014). Learning Through Noticing: Theory and Evidence from a Field Experiment. Quarterly Journal of Economics 129(3), 1311–1353.

Heal, G. and Millner, A. (2014). Uncertainty and Decision Making in Climate Change Economics. Review of Environmental Economics and Policy 8(1), 120–137.

Jondrow, J., Lovell, C.A.K., Materov, I.S. and Schmidt, P. (1982). On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model. Journal of Econometrics 19(2–3), 233–238.

Ma, X., Spielman, D.J., Nazli, H., Zambrano, P., Zaidi, F. and Kouser, S. (2014). Information Efficiency in a Lemons Market: Evidence from the Bt Cotton Seed Market in Pakistan. Selected Paper, Agricultural and Applied Economics Association Annual Meeting, Minneapolis, July 2014. International Food Policy Research Institute.

Mailier, P.J., Jolliffe, I.T. and Stephenson, D.B. (2006). Quality of Weather Forecasts: Review and Recommendations. Report to the Royal Meteorological Society.

Manski, C.F. (2004). Measuring Expectations. Econometrica 72(5), 1329–1376.

Meeusen, W. and van den Broeck, J. (1977). Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. International Economic Review 18(2), 435–444.

Michelson, H., Fairbairn, A., Maertens, A., Ellison, B. and Manyong, V. (2021). Misperceived Quality: Fertilizer in Tanzania. Journal of Development Economics 148, 102579.

Nelson, P. (1970). Information and Consumer Behavior. Journal of Political Economy 78(2), 311–329.

Nepal, M., Ashfaq, M., Sharma, B.R., Shrestha, M.S., Khadgi, V.R. and Bruno Soares, M. (2024). Impact of Weather and Climate Advisories on Agricultural Outcomes in Pakistan. Scientific Reports 14, 1036.

Norton, B.P. (2020). Ambiguity and Credence Quality: Implications for Technology Adoption.

MSc thesis, University of Illinois at Urbana-Champaign.

Open-Meteo (2025). Historical Forecast API: archived ICON model output. Weather data by

Open-Meteo.com (CC-BY 4.0); forecast model by Deutscher Wetterdienst.
Roudier, P., Muller, B., d'Aquino, P., Roncoli, C., Soumaré, M.A., Batté, L. and Sultan, B. (2014). The Role of Climate Forecasts in Smallholder Agriculture: Lessons from Participatory Research in Two Communities in Senegal. Climate Risk Management 2, 42–55.
Simonsohn, U., Simmons, J.P. and Nelson, L.D. (2020). Specification Curve Analysis. Nature Human Behaviour 4, 1208–1214.
Stigler, G.J. (1961). The Economics of Information. Journal of Political Economy 69(3), 213–225.
Suckall, N. and Bruno Soares, M. (2020). Valuing Climate Services: Socio-Economic Benefit Studies of Weather and Climate Services in South Asia. Asia Regional Resilience to a Changing Climate (ARRCC) Met Office Partnership. University of Leeds.
Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables. Econometrica 26(1), 24–36.
Vaughan, C. and Dessai, S. (2014). Climate Services for Society: Origins, Institutional Arrangements, and Design Elements for an Evaluation Framework. WIREs Climate Change 5(5), 587–603.