Journal of Hydrology 666 (2026) 134748

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Sociohydrological model structure deficiency assessment and hybrid model selection

Dennis Djohan[a], Julien Malard-Adam[b,c,d], Soham Adla[a], Saket Pande[a,*]

[a] *Department of Water Management, Delft University of Technology, Delft, the Netherlands*
[b] *G-Eau, Institut de recherche pour le développement (IRD), Université de Montpellier, Montpellier, France*
[c] விரிவாக்கக் கல்வி இயக்ககம், தமிழ்நாடு வேளாண்மைப் பல்கலைக்கழகம் *(Directorate of Extension Education, Tamil Nadu Agricultural University), Coimbatore, Tamil Nadu, India*
[d] *Institut français de Pondichéry (IFP), Puducherry, India*

HIGHLIGHTS

- Farm survey identifies sociohydrological (SH) model deficiencies in yield predictions.
- Mixed method of surveys and machine learning used to model SH structural errors.
- Hybrid SH model predicts better than SH only and machine learning only models.
- Structural errors arise due to absence of irrigation practices and farmer adaptations.

ARTICLE INFO

This manuscript was handled by Okke Batelaan, Executive Editor

Keywords:
Smallholders
Sociohydrology
Mixed methods
Machine learning
Yield predictions
Plot scale

ABSTRACT

Sociohydrology aims to deliver user-inspired solutions to water challenges, often through model-based understanding and simulation of local realities. However, sociohydrological modeling methodologies used to understand such complex human-water systems remain difficult to apply to many real-world case studies. Sociohydrological model predictions at daily to annual time scales of decision- making remain a challenge due to often difficult-to-acquire social sciences data, and missing or unknown feedbacks that lead to model structural errors, among other issues. This paper assesses and reduces model structural deficiencies of a smallholder sociohydrological (SH) model when applied to a case study of small-scale agricultural production in India, where variables from a farmer survey help alleviate structural deficiencies. A structural error model is proposed based on a regression model of nonlinear projection of the these variables to a Kernel space, called Kernel Principal Component Analysis (KPCA) based model. Based on this, a hybrid model that is a sum of the SH model and the structural error model is proposed. It offers significantly better yield predictions on 'unseen' (to the model) survey data than the SH only model. The hybrid model also performs better on yield prediction than a KPCA model alone, which predicts yields without any SH dynamics. This is because the hybrid model combines the structural error model that learns from the spatial pattern of observed yields with the temporal dynamics explained by the SH model alone. The results indicate that the structure of the SH model can be improved by further incorporation of irrigation and adaptive behaviour of farmers.

1. Introduction

Sociohydrology offers a human-centric perspective to examine the bidirectional feedbacks and nonlinear interactions between human and water systems (Di Baldassarre et al., 2025; Gu et al., 2025; Rachunok and Fletcher, 2023). Central to sociohydrological studies has been the interpretation of diverse phenomena that emerge from coupled water-human system dynamics such as the levee effect and demand supply dynamics through a variety of methods that include causal loop diagrams, surveys and system dynamic modelling (Kreibich et al., 2025; Pande and Sivapalan, 2017). The phenomena studied include the levee effect, the irrigation efficiency paradox, technology-mediated agricultural intensification and environmental Kuznets curves (Di Baldassarre et al., 2025). These studies have advanced conceptual models of water human interactions and novel mixed methods approaches to interpreting such phenomena. However, the impact of such studies in terms of providing operational guidance to design and implement interventions

[*] Corresponding author.
Email address: s.pande@tudelft.nl (S. Pande).

https://doi.org/10.1016/j.jhydrol.2025.134748
Received 26 May 2025; Received in revised form 5 December 2025; Accepted 6 December 2025

Available online 13 December 2025

that deliver the intended results has remained limited to few studies; see, for instance, Alam (2024); Amirkhani et al. (2022); Ghoreishi et al. (2021); Quesnel and Ajami (2017).

As an example, Alam (2024) recently proposed an agent based model interpretation of the phenomenon of supply-demand dynamics in the context of agricultural water interventions such as check dams in Gujarat, India, the proliferation of which was observed in secondary data such as groundwater data and agricultural census data. These mixed data sources suggested that check dams were constructed with the intention of recharging groundwater. However, they led to increased demand for water for irrigation due to the perception of farmers that the water supply had increased. This elucidated the need to incorporate feedbacks linked to human perceptions in models if these are to be used to design groundwater conservation policies implemented at the farm to catchment scale (Adla et al., 2025). Alam (2024) calibrated such feedbacks using data from surveys of farmer behaviour towards agricultural water interventions (Alam et al., 2022).

While sociohydrological models, due to the conceptual advancements they bring, are used to anticipate emergent dynamics and infer possibility space for solutions to avoid undesirable water futures, their predictions remain limited in their ability to predict at daily to annual scales and provide space for actionable interventions (Sivapalan and Blöschl, 2015; Srinivasan et al., 2025). For example, Roobavannan et al. (2017) and Li et al. (2025) developed sociohydrological models in agricultural and urban water contexts respectively, and although they were informed by available secondary social data, they were limited to the anticipation of trajectories at multidecadal to centennial scales. As a result, the solutions discussed were at annual or larger time scales, such as policies linked to reservoir operations in the Murrumbidgee River Basin and groundwater governance in Beijing, and needed more specific information to detail actions needed at scale. This has been widely acknowledged, that most sociohydrological models can only be used to 'anticipate' future dynamics in long term and not 'predict' at shorter time scales (Srinivasan et al., 2025). But predictability of sociohydrological models at decision making scales is critical to designing and implementing interventions at scale (Pande and Sivapalan, 2017).

One reason behind the challenge of predicting at actionable time scales is obtaining appropriate social data sets (Pande and Sivapalan, 2017; Srinivasan et al., 2025). This is primarily because these acivities are resource intensive (e.g., fieldwork to conduct social surveys) and can only be sampled at time resolutions coarser than the time scale of hydrological dynamics. Furthermore, which variables to collect information on depends on the aspects of water-human dynamics that may not be well understood. Not all feedbacks of water human dynamics may be known *a priori* (Kreibich et al., 2025), and this leads to model structural deficiency (Pande and Sivapalan, 2017). Approaches to identify missing feedbacks can include open-ended questions with practitioners (Haeffner, 2022) and machine learning or information theoretic models to find patterns in surveyed data that complement variables and data used in structurally deficient sociohydrological model simulations (Pande and Sivapalan, 2017).

The purpose of this paper is to propose a replicable methodology to assess deficiencies in such models, with a key focus on improving smallholder sociohydrological predictions at daily to annual scales. These scales are relevant for farmer decision making and for designing interventions such as good agricultural practices that can help alleviate farmer distress. As a case study, this paper investigates the structural deficiencies of a smallholder sociohydrological model of Pande and Savenije (2016) based on yields and other sociohydrological variables of cotton farmers observed in a social survey. Here, reliable prediction of yields every year at the smallholder plot scale is key to informing smallholder farmers of practices at daily to monthly scales to realize better yields and avoid financial distress (Mishra, 2006).

Predicting yields at the smallholding plot scale (e.g., at a 2 ha resolution) is challenging (Jabed and Azmi Murad, 2024). There are studies that use high to very high spatial resolution time series data of biophysical variables such as greenness, temperature and canopy cover, for instance, by using SkySat, PlanetScope and Sentinel-2 at 2 m to 20 m resolution data at various points in time during a growing season (Ponce-Pacheco et al. (2025) and references therein) to predict yields with good accuracy. However, there are no studies that provide forward-looking yield predictions to farmers and practitioners, such as agricultural extensionists, with an understanding of farm scale water-human dynamics, such as through the lens of smallholder sociohydrological model of Pande and Savenije (2016). This is needed to design and advise farmers on system-cognizant good agricultural practices so that they can act in a timely manner to improve their anticipated yields (Ayre et al., 2025).

Motivated by this, the study uses a simple machine learning model, based on a linear regression with principal components of explanatory variables in a nonlinear Kernel space (KPCA), to develop a predictive structural error model (see e.g., Honti et al. (2013); Pianosi and Raso (2012); Xu and Valocchi (2015) for similar error models). The difference between observed yields and those predicted by the model of Pande and Savenije (2016) is identified as possible structural deficiency in the model. While the sociohydrological model considers the temporal dynamics, the machine learning model learns the spatial patterns of observed yields unexplained by the sociohydrological dynamics as a function of various farm scale characteristics obtained from a social survey. Using the structural error model as the basis, this paper then presents the value that the structural error model adds in sociohydrological prediction in comparison to other model variants and discusses ways forward.

2. Materials and methods

2.1. Study area

India is one of the top producers of cotton in the world (Raghavendra and Reddy, 2020; Shwetha et al., 2022). Small and marginal farms (with 0–2 ha of land) account for more than 85 % of farm holdings and contribute to more than 50 % of India's agricultural output (Fan et al., 2013; Joshi and Tyagi, 2019). Nonetheless, the security of smallholder livelihoods continues to be challenged by climate change impacts, poor soil quality, price volatility, and a lack of access to assets, capital, technology, markets, and alternative employment opportunities (Fan et al., 2013; Joshi and Tyagi, 2019). Smallholder farmers are often forced to borrow money, not just for agriculture, but also for their daily needs (Sravanth and Sundaram, 2019). Central and state governments have passed several debt relief bills and implemented debt waiver schemes in the past to reduce the burden of debt (Kerala Legislative Assembly, 2012; Pathak and Chattopadhyay, 2021; RBI, 2009). However, multiple studies have shown that debt relief is not effective at combating the crisis in the long term and comes with detrimental effects on the borrowers' behaviour, such as no increase in investment or productivity, increasing default rates, and longer repayment periods (De and Tantri, 2014; Giné and Kanz, 2014; Kanz, 2012; Pathak and Chattopadhyay, 2021). Moreover, debt relief initiatives are mostly reactionary and may not solve the underlying causes of the crisis (Sravanth and Sundaram, 2019). This is exemplified by Maharashtra, which, despite being among the largest cotton producing Indian states, has considerably low cotton productivity and resource poor farms (Khan and Ansari, 2023; Pande and Savenije, 2016; Raghavendra and Reddy, 2020).

The study area focuses on such cotton producing districts in the north-eastern region (Vidarbha) of Maharashtra, namely Amravati, Wardha, and Yavatmal. The region is covered with black soil of volcanic origin (van Wirdum et al., 2019), which consists of 54 % clay, 32.5 % silt, and 13.5 % sand (Katti, 1979). This type of soil is suitable for growing cotton due to its high moisture retention capacity and the relatively high temperatures in the region (Janssen, 2020). Other crops grown here include cereals, pulses, oilseeds, and sugarcane (Kamble and Tikadar, 2024; Mishra, 2006; Pande and Savenije, 2016). The agronomic potential of this region is influenced by the rainy season during the summer months, with an average of 50–60 rainy days per year

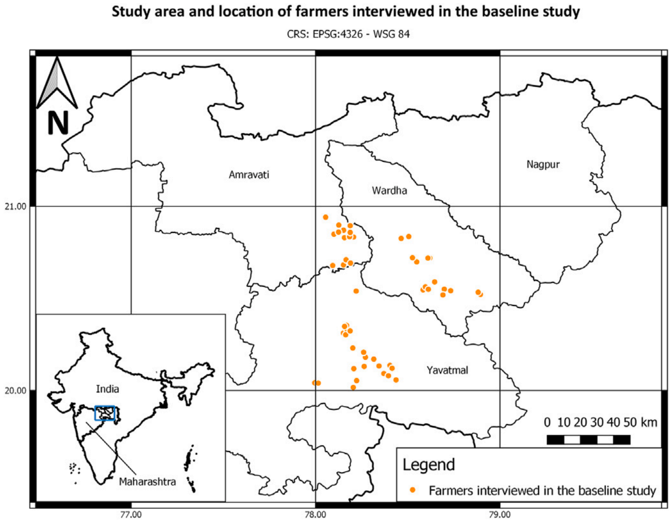

Fig. 1. Map of the study area showing the location of the 308 farmers interviewed (Hatch et al., 2022). Interviews were conducted in three districts of Maharashtra (India): Yavatmal, Amravati, and Wardha. These districts are located in the Vidarbha region in north eastern Maharashtra. Made in QGIS (QGIS Development Team, 2021).

Table 1
Details of the gridded data used. Source: Indian Meteorological Department (Pai et al., 2025; Srivastava et al., 2009).

Parameter	Unit	Spatial resolution	Temporal resolution (extent)
Precipitation (P)	mm/d	0.25 × 0.25°	daily (1975 - 2019)
Maximum temperature ($T_{\max}$)	°C	1 × 1°	daily (1975 - 2019)
Minimum temperature ($T_{\min}$)	°C	1 × 1°	daily (1975 - 2019)
Mean temperature (T)	°C	1 × 1°	daily (1975 - 2019)

Table 2
Description of the variables from the farmer survey (Hatch et al., 2022) and climate forcing used in the PCA, kernel PCA, and linear regression analysis.

Variables	Unit	Description
Family help	person(s)	Farm labour help from the family/children
Cotton area	ha	Total cotton area of the farmer
Seeds cost	INR/ha	Local cost of cotton seeds
Pesticide cost	INR/ha	Local cost of pesticide
Fertilizer cost	INR/kg	Local cost of fertilizer
Fertilizer amount	kg	Total fertilizer usage of the farmer
Soil depth	mm	Soil depth in the field
Latitude	°N	Latitudinal coordinate of the farmer
Longitude	°E	Longitudinal coordinate of the farmer
Precipitation	mm	Total precipitation in the 2018 planting season
ET_c	mm	Total reference evapotranspiration in the 2018 planting season
Irrigation	mm	Total irrigation in the 2018 planting season
Model yield	kg/ha	The predicted cotton yield per hectare using the SH model

and an annual rainfall of 150–200 mm (Ratna, 2012). Cotton cultivation requires between 700 and 1200 mm of water distributed over its 5–6 month growing period (Hussain et al., 2025; van Wirdum et al., 2019). Consequently, the absence of supplementary irrigation, coupled with other factors, can result in severe agricultural challenges, especially for smallholder farmers in the region (Pande and Savenije, 2016).

A baseline household survey of 308 farmers was conducted in 2019 (locations depicted in Fig. 1) (Hatch et al., 2022). The aim of the survey was to assess farmers' hydrological resources and socio-economic characteristics using information about farming practices and financial characteristics. The 308 farmers in the sample consist mostly of small to medium-sized farmers (0.5 to 10 ha); however, there are also several large farmers (>10 ha). Areas under cotton cultivation ranged between 0.5 ha and 35 ha, with an average of 4.7 ha (standard deviation SD = 4.1 ha). The average seed cotton yield was 1,500 kg/ha. 238 farmers were in debt (77 % of total farmers), out of which 107 farmers (45 % of the indebted farmers) had loans with interest rates higher than 10 %.

2.2. Data description

Gridded daily temperature and precipitation data are taken from the Indian Meteorological Department (IMD) (Pai et al., 2025; Srivastava et al., 2009), the details of which are given in Table 1. The mean daily temperature ranges from 15 °C to 37 °C. The reference evaporation (ETc) is estimated using the Hargreaves Equation from the temperature data (Hargreaves and Samani, 1985). For the water equivalent of extraterrestrial radiation (R_a [mm/d]), the average monthly values at 20° N latitude are used (Samani, 2000). It should be noted that the data resolutions are relatively low for representing plot-scale conditions.

The irrigation data are derived from a QGIS Soil and Water Assessment Tool (QSWAT) model of the study area (Dile et al., 2019; Janssen, 2020). QSWAT is a plug-in for the open source geographic information system QGIS. QSWAT was chosen as hydrological method to calculate the reservoir inflows at particular intake points since it is easy to use once set up, the data necessary to run the model are freely accessible, and the plug-in works via QGIS which creates visible results and is user friendly. Several data types are used to run the QSWAT plug-in: digital elevation map (DEM), land use map, soil type map and weather data for the specific location. The steps undertaken to simulate model reservoir inflows at particular points as part of the QSWAT plugin of QGIS include: 1) delineate a watershed using the DEM, and identify streams from the generated stream network flowing into intake points that lie within the watershed, 2) create the hydrological response units (HRUs). For this, the land cover and soil type data files are imported and HRUs classifications are created based on these datasets, and 3) finally, imported the weather data into the model. Once all data files are imported, clicking "ok" will implement the data in the model. Clicking on the option "Setup and Run SWAT Model Simulation" under the "SWAT Simulation" option will run the model; once the model is run, the SWAT Output can be generated and inflows at intake points for a period of study can be obtained. For more details on how QSWAT was set up and run for this study, readers are referred to Janssen (2020), p 21–24.

Variables from the survey (Hatch et al., 2022) and climate data (Table 1) are used to quantify the total error of the model. An overview of these variables is shown in Table 2.

2.3. Smallholder sociohydrological (SH) model

The sociohydrological (SH) model used is based on the model by Pande and Savenije (2016) as shown in Fig. 2. The model provides insight into the system dynamics of smallholder farmers through six assets, namely water storage capacity, capital, livestock, soil fertility, grazing access, and labour. This study focuses on the climate variability, soil moisture, and crop production parts of the model. In addition, the labour availability factor is adjusted in the calibration phase due to its direct influence on the crop production.

2.4. SH model calibration and diagnosis

Changes in the SH model structure are first implemented based on known deficiencies identified by Pande and Savenije (2016) and then calibrated. The model's performance is evaluated by comparing the simulated yields with the observed yields from the survey (Hatch et al.,

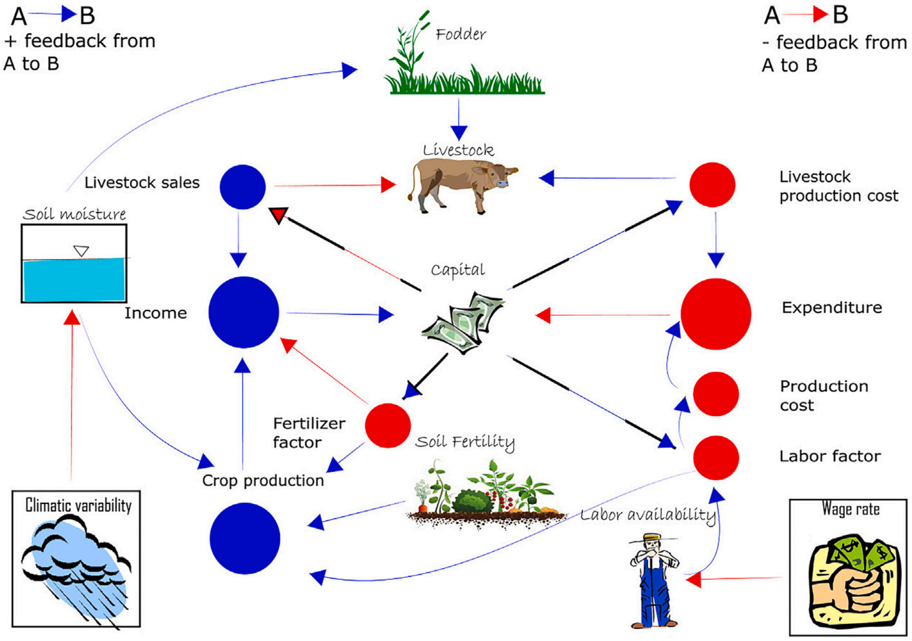

Fig. 2. The base SH model obtained from Pande and Savenije (2016). The model captures the dynamics of a smallholder farmer through their assets (state variables). These include water storage, grazing area, livestock, capital, soil fertility, and labour. This research focuses on the following parts of the SH model: climate variability, soil moisture, and crop production.

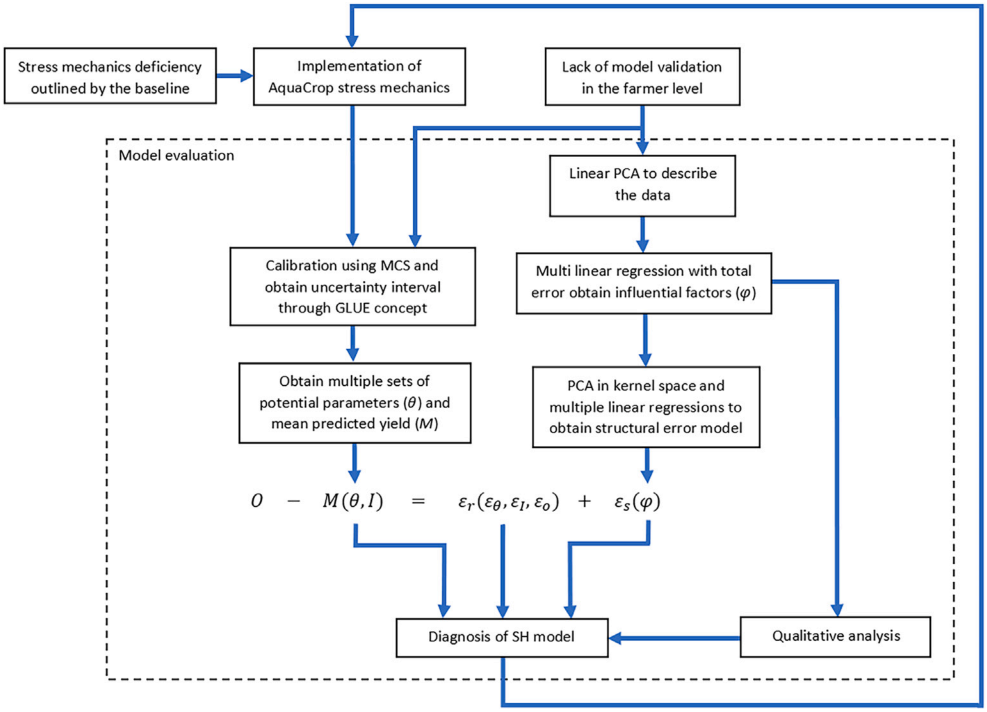
Stress mechanics deficiency
outlined by the baseline
Implementation of
AquaCrop stress mechanics
Lack of model validation
in the farmer level
Model evaluation
Linear PCA to describe
the data
Calibration using MCS and
obtain uncertainty interval
through GLUE concept
Multi linear regression with total
error obtain influential factors (φ)
Obtain multiple sets of
potential parameters (θ) and
mean predicted yield (M)
PCA in kernel space and
multiple linear regressions to
obtain structural error model
$O - M(\theta, I) = \varepsilon_r(\varepsilon_\theta, \varepsilon_I, \varepsilon_o) + \varepsilon_s(\varphi)$
Diagnosis of SH model
Qualitative analysis

Fig. 3. SH model calibration and diagnosis. The AquaCrop temperature and water stress mechanics are implemented in Pande and Savenije (2016) and KPCA based structural error model added to diagnose and ameliorate SH model structural deficiency.

2022). The total error is quantified, and the variables from Table 2 that might be associated with the remaining unknown deficiencies in the total error are identified. Fig. 3 provides an overview of these steps.

The model of Pande and Savenije (2016) does not take into account the stress caused by water deficit on crop yield. Temperature stress has also not been considered. These are addressed by implementing the daily stress mechanics of the Food and Agriculture Organization's (FAO) AquaCrop model (Allen et al., 1998; FAO, 2012) and are provided in equations (9,16) and (26)–(28) respectively in Supplementary Material Section 2.

The parameters of the updated SH model are then calibrated using Monte Carlo Sampling (MCS). This achieves two purposes. First, MCS serves as a sensitivity analysis to check which parameters the performance of the model is sensitive to. When a parameter is deemed not influential (i.e., similar performance scores across its value range), it is removed from the calibration to avoid adding unnecessary complexity to the model. This also limits the number of parameters that are used for the calibration and reduces the number of iterations needed to find the optimal parameter sets. Secondly, once the sensitive parameters are found, multiple sets of optimal parameter values (θ) are obtained. This follows the concept of Generalized Likelihood Uncertainty Estimation (GLUE), where different sets of parameters are considered equally likely as a simulator of the system (Beven and Binley, 1992). The behavioural parameter sets (approximately top 10 % in performance) are used to estimate the uncertainty in the model predictions.

The SH model is run from the year 2010 to 2018 using the optimal parameter values. The modelled yield ($M(\theta, I)$) in 2018 is then compared with the observed yields (O), obtained from the 2019 farmer survey (Hatch et al., 2022) to obtain the total error. To diagnose the model, an additive structural error model based on Kennedy and O'Hagan (2001) is used, in which the total error is considered to be made up of two parts. The first is the residual error (ε_r) which is a function of parameter errors (ε_θ), input errors (ε_I), and observation errors (ε_o). The second is defined as structural error (ε_s), which is a function of variables that are influential and correlated with the structural error (φ). These variables φ are obtained from the survey of Hatch et al. (2022) and climatic conditions and are shown in Table 2. The structural error model is obtained by performing Principal Component Analysis (PCA) in a higher-dimensional nonlinear space, called Kernel space, of the influential variables. Here complex nonlinear relationships become linear (KPCA) and therefore linear regressions (between total error and influential variables) can be applied in this space. This is called a Kernel Principal Component Analysis (KPCA) structural error model, similar to Support Vector Machines that perform a linear regression in Kernel space (Schölkopf et al., 1998). When observed yield, instead of total error, is used as dependent variable in a KPCA based model, a KPCA only model is obtained.

In order to assess SH model structure deficiency and model improvement, three different model variants are thus compared: SH model only (with no structural error modeling), KPCA only (KPCA model of observed yields, not modeling structural error) and SH + KPCA structural error model (hybrid, sum of SH only modelled yield + KPCA structural error model). The KPCA structural error model linearly adds to the SH only model prediction to provide the SH + KPCA structural error model variant without any bidirectional coupling.

2.5. Evaluation of model variants

Four objective functions are used to evaluate the performance of the different models. The first metric is the Nash-Sutcliffe (NS) coefficient; see Eq. (1) (Nash and Sutcliffe, 1970), where Y_i and Y_i' are the observed and modelled yields, respectively, of the ith farmer. The NS coefficient of the log values of Y and Y' (NS_{log}) is also used; see Eq. (2). Another metric used for performance evaluation is the Mean Absolute Error (MAE) - see Eq. (3). Lastly, the coefficient of determination r^2 value is also used. Eq. (4) shows the calculation of r^2, where $\hat{Y}_i$ is the corresponding y-value of a linear regression between model predictions and observed yields. $\bar{Y}$ is the mean of observed yields. Supplementary Material Section 3 provides more details on the error typology, its description, calibration and how uncertainty intervals are derived.

$$NS = 1 - \frac{\sum_{i=1}^{n} (Y_i' - Y_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \tag{1}$$

$$NS_{log} = 1 - \frac{\sum_{i=1}^{n} (log(Y_i') - log(Y_i))^2}{\sum_{i=1}^{n} (log(Y_i) - log(\bar{Y}))^2} \tag{2}$$

Table 3
Various kernel functions used in the study. The Kernel parameters γ and c are estimated on the training data.

	$K(\mathbf{x}_i, \mathbf{x})$
RBF	$\exp(-\gamma \parallel \mathbf{x}_i - \mathbf{x} \parallel^2)$
Sigmoid	$\tanh(\gamma \mathbf{x}_i \cdot \mathbf{x} + c)$
Poly deg d	$\parallel \gamma \mathbf{x}_i \cdot \mathbf{x} + c \parallel^d$
Cosine	$\frac{\mathbf{x}_i \cdot \mathbf{x}}{\lVert\mathbf{x}_i\rVert \lVert\mathbf{x}\rVert}$

$$MAE = \frac{\sum_{i=1}^{n} |Y_i' - Y_i|}{n} \tag{3}$$

$$r^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \tag{4}$$

2.5.1. KPCA based models

While the components obtained from linear PCA can be useful to evaluate the first order effects of the variables in Table 2 on SH model structural deficiency, this may be insufficient to capture the underlying non-linearities in the data. Thus, the model variables are mapped to a higher dimensional space using kernel functions where they are linearly separable (Raschka and Mirjalili, 2019; Schölkopf et al., 1998). PCA is performed in this new kernel space and a linear regression with selected principal components in the kernel space is performed with total error as the dependent variable to obtain a KPCA based structural error model. When the dependent variable is the observed yield, the 'KPCA only' model is obtained.

A KPCA based model is obtained as follows. For a given set of independent variables $\mathbf{x}$ and corresponding data $\{\mathbf{x}_i, i = 1, .., n\}$, a nonlinear mapping $\varphi(\mathbf{x})$ (a $M \times 1$ vector) is applied to project the data into a higher $M \leq \infty$ dimensional space. KPCA is then used to find top k principal components out of a maximum of M principal components. The principal components are expressed in terms of a Kernel function, $K(\mathbf{x}_i, \mathbf{x}) = \varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}) = \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x})$. Different kernels are used, as shown in Table 3. A jth principal component ($j \leq M$) is given by,

$$\beta_j(\mathbf{x}) = \frac{1}{\sqrt{n\lambda_j}} \sum_{i=1}^{n} \alpha_i^j K(\mathbf{x}_i, \mathbf{x})$$

Here λ_j are eigenvalues and $\boldsymbol{\alpha}^j = \{\alpha_i^j : i = 1, ..n\}$ are the corresponding eigenvectors, obtained by solving the eigenvalue problem: $n\lambda\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{\alpha}$. Here, $\mathbf{K}$ is the symmetric Gram matrix of size $n \times n$ with elements defined as $K_{i,j} = \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$. When the Kernel function is chosen as a simple dot product, $K(\mathbf{x}_i, \mathbf{x}) = \mathbf{x}_i \cdot \mathbf{x}$, then the nonlinear mapping reduces to a linear mapping, where $\varphi(\mathbf{x}) = \mathbf{x}$ and obtained principal components correspond to linear principal components (i.e., PCA).

Finally, the KPCA based model is applied using the top k nonlinear principal components $\{\beta_j(\mathbf{x}), j = 1, .., k \leq M\}$ as predictors in a linear regression model with y: $y = \sum_{j=1}^{k} w_j \beta_j(\mathbf{x}) + b$. Here b is a bias term and $\mathbf{w} = \{w_j, j = 1, .., M\}$ is the set of linear regression coefficients in the non-linear space, estimated using least squares regression. Its estimate, $\hat{\mathbf{w}}$, is given by: $\hat{\mathbf{w}} = \tilde{\Lambda}^{-1}\mathbf{B}^T \boldsymbol{y}$, where $\boldsymbol{y}$ is an $n \times 1$ column vector of the observations of the dependent variable, $\tilde{\Lambda} = n\mathrm{diag}(\lambda_1, .., \lambda_M)$ is n times the diagonal matrix of the eigenvalues $(\lambda_1, .., \lambda_M)$. $\mathbf{B} = \boldsymbol{\varphi}\mathbf{V}$ is the feature matrix in nonlinear space, where $\boldsymbol{\varphi}$ is a $n \times M$ matrix containing transformed data points. For example, its ith row is the $1 \times M$ row vector representing nonlinear transformation of $\mathbf{x}_i$, given by $\varphi(\mathbf{x}_i)^T$, and $\mathbf{V}$ is a $M \times M$ matrix with its j^{th} column given by $\mathbf{v}^j = \sum_{j=1}^{n} \alpha_i^j \varphi(\mathbf{x}_i)$.

The choice of dependent variable y determines the type of KPCA based model. When y is the observed yield, the model that is obtained is the 'KPCA only' model and, when y is the total error (i.e., differences between observed yields and SH model predictions), the KPCA structural error model is obtained.

The top k principal components in the nonlinear space are obtained such that the proportion of variance explained by the top k principal components is at least (≤ 90 %) of the total variance: $\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{M} \lambda_j} \leq 90$ %.

The KPCA command in the scikit-learn library of the Python programing language is used (Pedregosa et al., 2011) to implement the KPCA algorithm. Outliers, defined as data points with a z-score (obtained by subtracting the mean and dividing by the standard deviation) greater than 3.5, are excluded from the analysis. Cross validation is performed in order to select the best kernel function that describes the data and to prevent overfitting from a collection of sigmoid, radial basis function (RBF), cosine, and polynomial (degree 2 to 5) kernels (Table 3). The available data are split into 75 % training data and 25 % test data, so that there is sufficient data to also calibrate the Kernel parameters (See Table 3). Using the training data, the kernel functions are used to obtain various models. Each model is tested using the test data to predict the observed output variable of interest. The prediction is compared with the observation to find the best performing kernel and model. The performance is scored based on MAE, NS, and r^2.

2.5.2. Residual error

The residuals between the prediction of the KPCA based structural error model and the observed total error are attributed to the residual error. The distribution of the residual errors is deemed to be caused by uncertainties in the input, parameter, and observational values.

3. Results

The overall performance of the crop growth components of the SH only model is described in this section, followed by the results of the structural error model in the kernel space and a comparison of the performances of model variants: SH only model, KPCA only model and hybrid model (SH only + KPCA structural error model).

3.1. SH only model

The seasonal evolution of soil moisture, water stress factor (Ks), and canopy cover (CC), for a farmer in the study area in 2018 is shown in Fig. 4. The evolution of soil moisture closely follows the rainfall events and is inversely related to the transpiration. Similarly, the water stress factor follows the changes in soil moisture. There is no water stress at the start of the planting season; thus, the canopy cover growth is non-stressed. As the soil moisture drops, the canopy development starts to experience senescence. It can be observed that transpiration slowly drops to 0 as the soil moisture reaches the wilting point.

3.1.1. Crop yield prediction

After calibration, the average crop yields and their uncertainty intervals are calculated using behavioural parameter sets (θ) based on thresholds for the four objective functions used to evaluate model performance. The threshold values used in this experiment are: $NS \geq -1.0$, $NS_{log} \geq -2.0$, $MAE \leq 600$ kg/ha/yr, and $r^2 \geq 0.003$. We follow Generalized Likelihood Uncertainty Estimation (GLUE) approach of Beven and Binley (1992) where thresholds are a subjective choice to identify what a modeler thinks are behavioural parameters in their handling of associated residuals. Readers are referred to Beven et al. (2012) for a further discussion on subjective choices of thresholds. Based on these thresholds a total of 1,557 parameter sets out of 10,000 simulations were used to obtain the uncertainty intervals (approximately top 10 %). 54.2 % of the 308 farmers fall within the uncertainty intervals. Fig. 5 shows the plot of calculated yields against the observed yields for all the farmers of the study area.

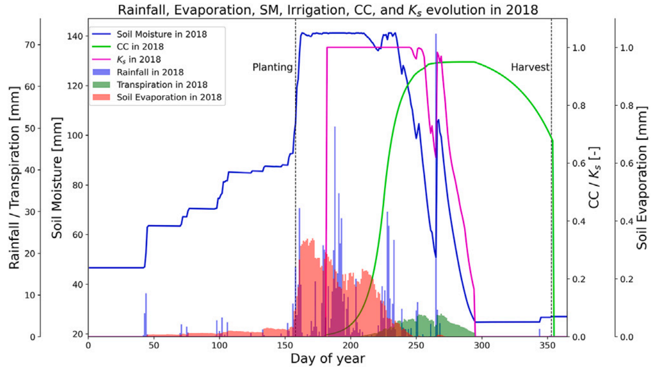

Fig. 4. Seasonal evolution of soil moisture, water stress factor (Ks), and canopy cover (CC) due to rainfall, transpiration, and soil evaporation. Soil moisture increases and decreases during the rainy season and at the peak of the planting season, respectively. K_s starts to decrease with depleting root zone soil moisture. The observed CC grows rapidly during the time when water is abundant, and senescence is triggered when water is depleted.

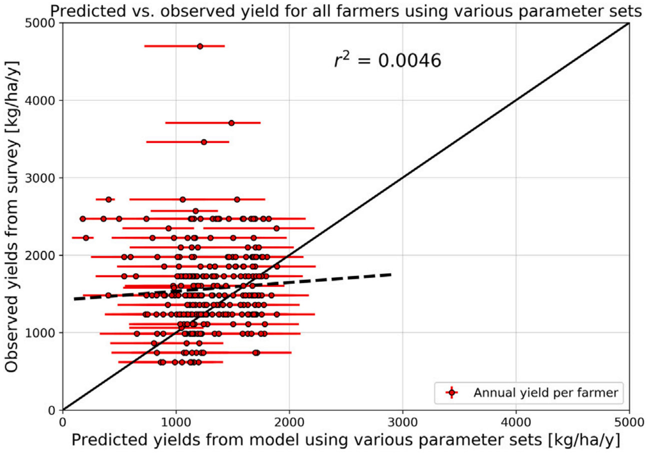

Fig. 5. Yield comparison between the mean predicted yield by SH model only and observed yields. The r^2 value is low at 0.0046 indicating that the model results are not correlated to the observed data.

Table 4
Ranges of calibrated parameters and objective functions of SH model only. The poor values for the objective functions indicate poor performance of the model.

	HI [-]	t_{cco} [days]	$f_{shape,lab}$ [-]	NS [-]	NS_{log} [-]	MAE [kg/ha]	r^2 [-]
Max	0.4	7	−6.5	−0	−0.7	553	0.008
Min	0.3	30	−10.0	−1.0	−1.9	447	0.003

The three parameters selected for calibration are HI, t_{cco}, and $f_{shape/lab}$ (harvest index, time to reach initial canopy cover in days and shape of the labour factor function; all are defined in Supplementary Material Section 1 and explained in Supplementary Material Section 2). HI is influential due to its direct influence on yield, i.e., the various parameters calculated during the crop production can be compensated by adjusting the value of HI. Similarly, $f_{shape/lab}$ has a direct influence and can compensate for the uncertainties in labour availability. Lastly, t_{cco} is influential as it determines whether the canopy cover will reach its maximum before the rainfall period ends. Table 4 shows the ranges of objective function scores obtained from the optimal parameter sets.

All of the objective function scores are poor, as is also evident from Fig. 5. The negative NS and NS_{log} values suggest that the model performs worse than using the average observed yield as a prediction. The MAE represents around 30 % deviation on average of the predicted yields from the observed yields. Lastly, according to the r^2 values, the predicted yields do not seem to be correlated with the observed yields. More details on the variation of objective functions with the parameter

Table 5
Rotated component matrix; the bold numbers indicate significant loadings attributing corresponding variables to the components. The correlation condition for a variable to be significant is either ≥ 0.5 or ≤ −0.5.

	Principal Component				
	1	2	3	4	5
Family help	−0.092	−0.132	−0.073	−0.083	**0.896**
Cotton area	0.102	0.096	**−0.852**	−0.024	0.083
Seeds cost	0.088	−0.110	0.025	−0.069	−0.188
Pesticide cost	0.114	−0.105	0.008	**0.823**	−0.018
Fertilizer cost	−0.069	0.068	0.010	**0.843**	0.091
Fertilizer amount	0.132	0.433	0.152	0.165	0.411
Soil depth	**0.578**	−0.492	0.434	0.015	−0.032
Latitude	**0.913**	−0.118	0.213	0.001	−0.169
Longitude	0.027	**0.905**	−0.117	−0.043	−0.077
Precipitation	−0.245	**0.819**	−0.148	−0.071	0.113
Evaporation	**−0.910**	−0.016	−0.198	0.032	0.153
Irrigation	**−0.797**	0.040	0.275	−0.098	−0.056
SH only yield	0.434	−0.122	**0.750**	−0.006	0.039

values are given in Supplementary Material Section 4. These poor results of the SH only model provide motivation for structural error modelling in order to ameliorate model errors.

3.2. Structural error model in linear space

3.2.1. Linear principal component analysis

The 13 variables that are tested against the total error are listed in Table 2. These variables are reduced to five principal components (PCs). The five PCs explain for 70.7 % of the total variance of the 13 variables. This value, being greater than 60 %, is deemed acceptable as the threshold to select the number of principal components (Hair, 2009). Table 5 gives the list of 13 variables tested, the reduced principal components (PC 1–5) and the correlation of latter against each of the 13 variables. Factor loading values of either ≥ 0.5 or ≤ −0.5 are considered to be significant and are used to associate corresponding variables to each PC (Hair, 2009). The significant variables for each of the five PCs can be seen in bold in Table 5. Therefore PC1 and PC2 can be perceived as the soil depth, location, and hydrology specific components, PC3 is related to cotton area and modelled yield, PC4 is associated with input costs, and PC5 is linked with available family help.

3.2.2. Linear regression of the principal components with the total error

The linear regression results with the five principal components as independent variables and the total error as the dependent variable are shown in Fig. 6. Note that the components that are not correlated may still influence the total error in a nonlinear space such as a kernel space (see Section 3.3), which motivated the use of KPCA analysis. The first principal component (PC1) which represents the variation of soil depth, evaporation, irrigation, and latitude, is not correlated with the total error. PC2 shows a weak correlation and tends to overestimate at higher precipitation and longitude values.

PC3 has the strongest correlation compared to the other PCs. From this, it seems that the SH only modelled yield does not capture the effects of the crop grown area very well. This is because PC3 with strong influence of cotton area is correlated with the total error of observed yields not explained by the SH only model. PC4 also has no correlation. The costs and prices do not impact the total error. The SH only model assumes that even in negative capital conditions, farmers can proceed to the following year by cutting expenditure on certain activities - see Pande and Savenije (2016) for details on the socioeconomic aspect of the model. Further, the change in capital does not affect the seed and fertilizer available to the farmer, which implies that the yield does not change and total error stays the same. PC5 represents family help, and it has a weak correlation. This is consistent with the findings of the sensitivity analysis of the labour factor shape. Eq. (5) shows the linear model using the significant components PC2 and PC3. Combined, they account for 16.2 % of the variance of the total error. This further motivates the regressions of total error and the independent variables (see Table 5) in nonlinear kernel space using KPCA, the assumption being that nonlinear transformation of independent variables may explain a greater percentage of the variability of the dependent variable.

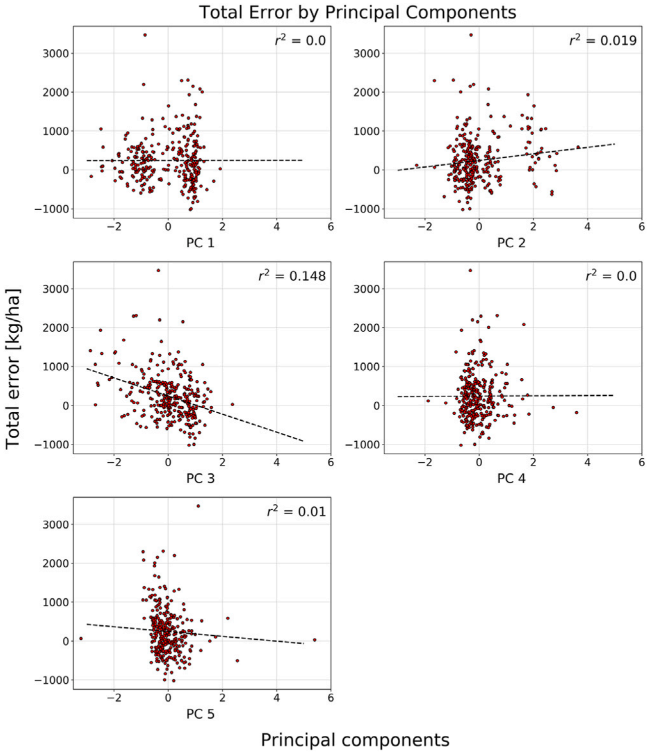

Fig. 6. Linear regressions between the linear principal components (from PCA) with the total error. PC 1 is associated with evaporation, irrigation, soil depth, and latitude. PC 2 is associated with precipitation and longitude. PC 3 is associated with the cotton area and SH model yield. PC 4 is associated with pesticide and fertilizer costs. PC 5 is associated with family help.

$$\varepsilon_{s,linear} = -232PC3 + 84PC2 + 241 \tag{5}$$

3.3. Structural error model in kernel space

3.3.1. Kernel choice and comparison of the performances of SH only, KPCA only and hybrid model variants

Table 6 provides an overview of the performance of various kernel models on both training and test datasets. There are significant differences in performance between the kernel functions. However, the NS score tends to decrease for all kernels between the training and test data, which may indicate overfitting. This was particularly evident for the cosine, RBF, and sigmoid kernels. The kernel with the best test data performance, the second-degree polynomial, was selected for further analyses. Fig. 7 illustrates the comparison of the prediction of the total error using the regression model, i.e., the structural error model, and the observed total error. The structural error model is able to capture approximately 34.6 % ($r^2 = 0.346$) of the variance of the total error. The difference between the predicted and observed error is attributed to the residual error (ϵ_r) and its distribution can be seen in Fig. 8. Assuming a Gaussian distribution, it has a standard deviation of $\sigma = 429$ kg/ha. The residual errors may be caused by the effects of structural deficiencies not explained by the model and their distribution is used to quantify the uncertainty interval of the yield prediction.

Table 6
Performance comparison of various kernel functions on training and test data. The kernels evaluated are Polynomials (degree 2 to 5), Cosine, Radial Basis Function (RBF), and Sigmoid. The kernels are evaluated using Mean Absolute Error (MAE) and Nash-Sutcliffe value (NS).

	Train Data		Test Data		All Data	
	MAE [$\frac{kg}{ha}$]	NS [-]	MAE [$\frac{kg}{ha}$]	NS [-]	MAE [$\frac{kg}{ha}$]	NS [-]
RBF	292	0.436	413	0.147	322	0.352
Sigmoid	349	0.292	396	0.0736	361	0.228
Poly deg 2	**329**	**0.36**	**371**	**0.309**	**340**	**0.345**
Poly deg 3	318	0.402	404	0.172	339	0.335
Poly deg 4	362	0.239	400	0.133	372	0.208
Poly deg 5	386	0.139	407	0.0686	392	0.119
Cosine	344	0.311	406	0.121	359	0.256

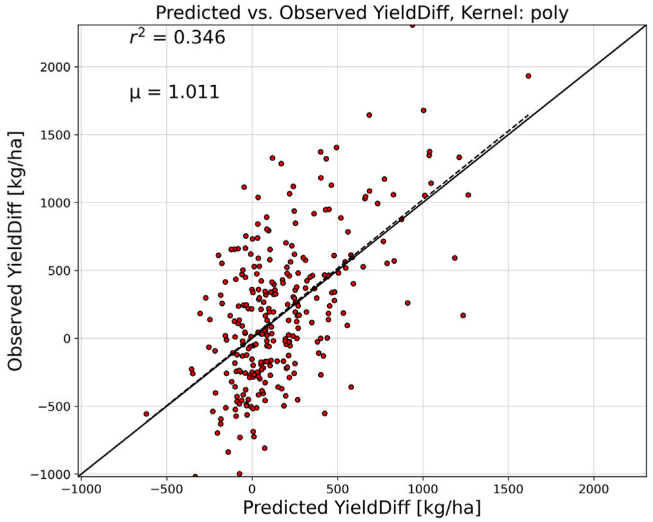

Fig. 7. Prediction of total error (defined as YieldDiff) using the KPCA structural error model. The structural error model is able to capture around 34.6 % of the variance of the total error. Also shown is the best fit line (dashed line) that has a slope, μ, of close to 1, indicating low bias.

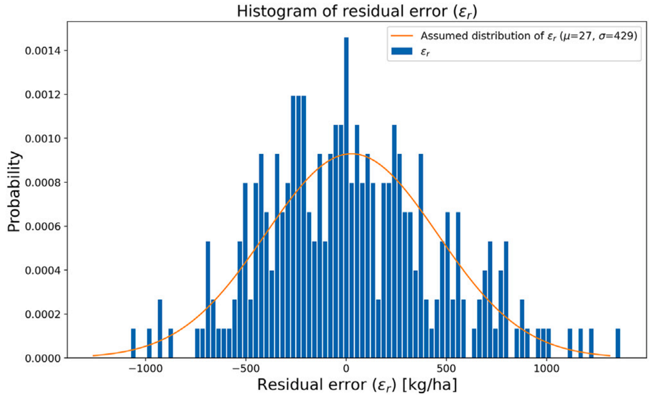

Fig. 8. Distribution of ε_r and its assumed Gaussian distribution (with σ = 429 [kg/ha] and mean μ = 27 [kg/ha]). This is assumed to be the uncertainty interval of the predictive model.

3.3.2. Adjusted prediction of cotton yield

The calculated yield from the SH only model is adjusted with the structural error model, which provides a better prediction than that from the SH only model. Table 7 shows the comparison between the SH only model, the combined hybrid model (SH + structural error model) and the KPCA only model. Note that KPCA only model is obtained when the dependent variable in the KPCA model regressions is chosen to be the observed yield instead of the total error. When the dependent variable is the total error, the structural error model is obtained, which when added to SH gives the SH + structural error model. The adjustment using the structural error model improves the prediction in all four scores. Also, the hybrid model (SH + structural error model) is better than the individual SH and KPCA only models. This improvement can also be seen in Fig. 9. The predicted yield is better compared to the predicted yield shown in Fig. 5. The uncertainty interval based on the assumed Gaussian distribution of ε_r can be seen as the yellow interval around the fitted line of the predicted yield. The regression line falls close to the $y = x$ line.

Table 7
Comparison of model scores between the SH only model, the hybrid (SH + structural error model) model, and the KPCA only model, showing an important improvement in all four scores.

	MAE [kg/ha]	NS [-]	NS_{log} [-]	r^2 [-]
SH only model	489	−0.624	−1.145	0.005
SH + structural error model	**371**	**0.1917**	**0.2578**	**0.201**
KPCA only	393	0.1263	0.1882	0.137

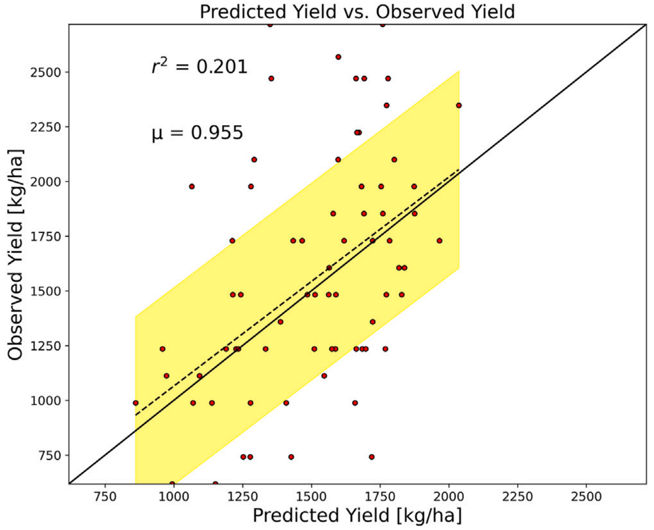

Fig. 9. Predicted vs. observed cotton yield by the SH + structural error model (hybrid model). The yellow area surrounding the best-fit line is the uncertainty interval obtained from the distribution of ε_r. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The addition of the structural error model improves the yield prediction NS score from −0.624 to 0.192 and r^2 from 0.005 to 0.201. This improvement indicates that the structural error model can partially compensate for the structural errors in the SH model, highlighting the presence of structural error in the SH model. The fact that the hybrid model outperformed the KPCA only model (NS = 0.126) also suggests that the SH model contributes important sociohydrological processes that cannot be replicated by a purely statistical KPCA model.

4. Discussion

The results of the sociohydrological (SH) model showed that the total error in modelling observed yields is large and indicates that structural errors are prevalent. The SH model by itself does not perform very well. However, when used in combination with the structural error model, the crop yield prediction scores improved to NS = 0.192, NS_{log} = 0.258, MAE = 371 kg/ha and r^2 value of 0.201, which represents better results than those obtained by either SH only or KPCA only models. This

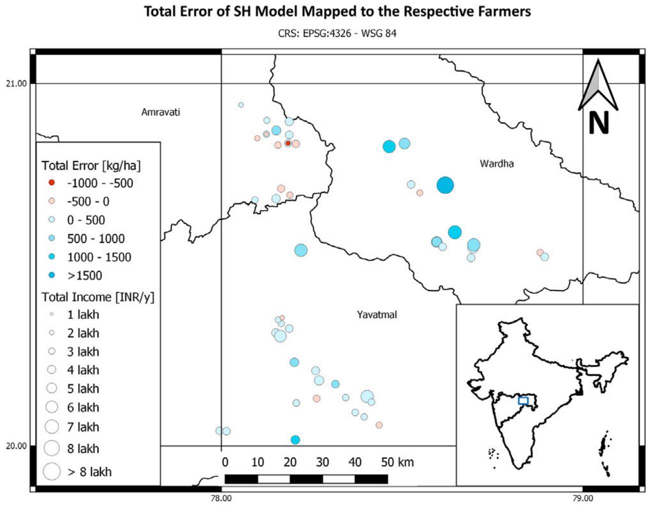

Fig. 10. Map of the mean total error and total income of the farmers in each village. Made in QGIS (QGIS Development Team, 2021).

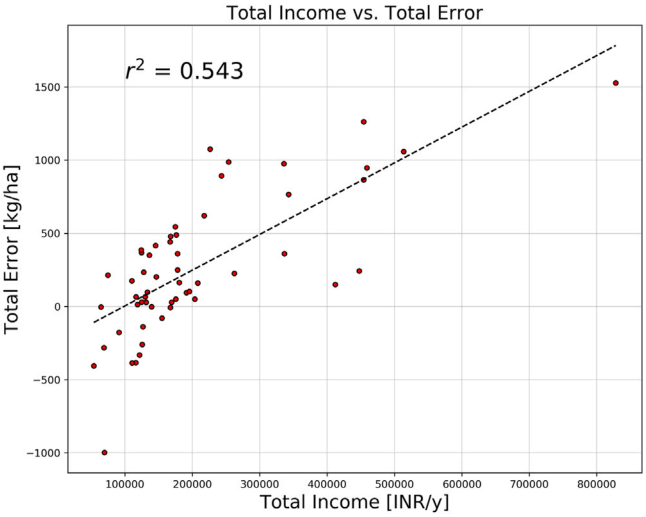

Fig. 11. Scatter plot of the mean total income vs. total error across villages ($\epsilon_r + \epsilon_s$).

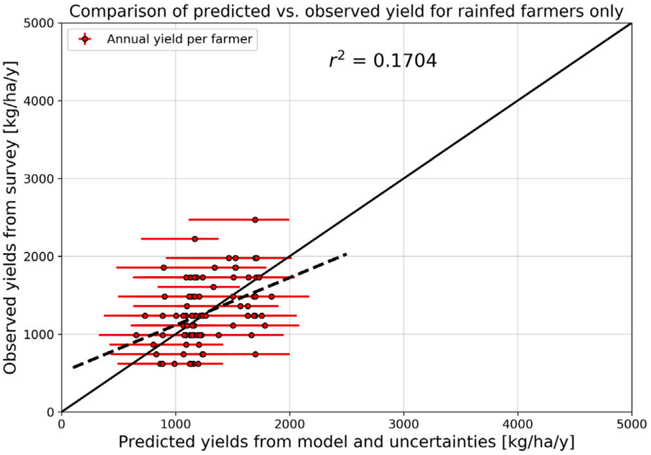

Fig. 12. Observed vs. SH model only predicted cotton yields of rainfed farms.

indicates that the poor performance of the SH model alone is likely caused by the prevalence of structural errors in the model.

The hybrid model selection approach to model smallholder sociohydrological systems is novel. The novelty lies in its mixed methods approach of using household surveys and a crop growth-water balance model to develop the structural error model. Both the mixed method approach to building a structural error model and demonstrating that it improves sociohydrological model prediction at scales of smallholder decision making, e.g., of 2 ha and daily time steps, is new. Increased accuracy of sociohydrological models at scale also means that interventions can now be designed to achieve intended outcomes with more confidence. This is a step forward when compared to existing sociohydrological models of diverse water human systems, both in terms of methods used and predictability, that mostly have remained conceptual and offer policy insights at coarse scales (Kreibich et al., 2025).

Given the system cognizance and improved accuracy of the hybrid model, it has been operationalized to interact with farmers and provide actionable advice on the choice of crops and irrigation (Adla et al., 2025; Ponce-Pacheco et al., 2025). For this the hybrid model has been web-enabled, in the form of an app called Makara (Ponce-Pacheco et al., 2025), so that farmers can themselves run the model for their farm system specifications and receive advice at plot and daily scales. Similar advice is also available to practitioners such as agricultural extensionists so that both extensionists and farmers can act on it to improve smallholder livelihoods. This is in contrast to coarse policy suggestions, such as the provision of alternative income sources to farmers, that the SH model could only offer due to limited predictive capacity of the SH model in Pande and Savenije (2016). Nonetheless, the accuracy of predictions of the hybrid model remains modest. Ways forward to improve the accuracy include (1) ingesting higher-resolution spaceborne datasets of biomass production and soil moisture than those that are currently being used, (2) considering additional socioeconomic factors such as those linked to prevalent cultural practices, and (3) using multiple machine learning algorithms to investigate if the accuracy of the structural error model predictions can be improved (Ponce-Pacheco et al., 2025).

The linear PCA and the multiple linear regressions indicated that precipitation, longitude, crop area, predicted yield, and family help are correlated with the total error. Among these, crop area and predicted yield have the strongest correlation. A selection of such variables, data for which were collected during a survey (Hatch et al., 2022), was then used in the KPCA structural error model. While the model aided in improving model prediction, it remains an algorithm that does not help in improving our understanding of farm system dynamics. Though this is acceptable when its purpose is prediction, more effort is needed to unpack such algorithms if the understanding of missing or new feedbacks is to be improved. This also relates to how new data and variables can be discovered and incorporated into a sociohydrological model. For example, Fig. 10, which shows the mean total error and total income of the farmers across villages, are correlated ($r^2 = 0.54$). Note that total income was not used in the structural error model. The self reporting of total income is often biased and in general measured through indirect indicators such as the type of house and the number of durables such as a television, refrigerator, motorbike, or other possession. The strong correlation between mean income and total error is more evident when plotted as scatter in Fig. 11 and suggests that richer farmers have better ways to mitigate water scarcity, more capital to afford fertilizer, and access to better tools to safeguard their yields and income. As another example, Fig. 12 shows stronger correlation between observed versus SH only model predicted yields when including only rainfed farmers ($r^2 = 0.17$) than when both irrigated and rainfed farmers are considered ($r^2 = 0.0046$, see Fig. 5). This significant difference in the performance indicates that irrigation mechanics should be investigated further. Exploratory interviews or workshops can be conducted with stakeholders to explore differences in behavioural dynamics between irrigated and rainfed farmers or regarding farmers with high levels of capital. These insights can then be used in a survey to obtain quantifiable variables and implement appropriate behavioural rules in the model. Translating observed structural deficiencies in this manner into feedbacks can also be supported by techniques such as symbolic regression to synthesize feedbacks (Munoz et al., 2025). The relevance of such feedbacks to the sociohydrological dynamics can then be tested as hypotheses by inserting them into the model and testing whether they improve model predictions at scale. This will be a transformative step that introduces missing dynamics based on the learnings of the structural error model.

5. Conclusion

Sociohydrological (SH) studies have advanced our understanding and interpretation of diverse phenomena that emerge from coupled water human system dynamics. Yet the impact of such studies in terms of providing operational guidance has remained limited to a few studies. This is because such models have traditionally been used to anticipate future system dynamics and are not sufficiently skilled in predictions at the plot and daily to annual time scales due to a lack of data and inadequate system understanding. Taking a smallholder sociohydrological system in eastern Maharashtra India as a case study, this paper presents a mixed methods machine learning approach that deploys both quantitative and qualitative models and data to improve predictions and system understanding of such models.

Ascribing structural deficiency, due to limited system understanding, to the difference between observed yields and those predicted by a smallholder SH model, the paper implemented Kernel Principal Component Analysis (KPCA) based regression model and developed a predictive structural error model. Based on a farmer survey on various aspects of their activities, this model learned the spatial patterns of observed yields unexplained by the sociohydrological temporal dynamics. A hybrid model as the sum of the structural error model and the SH model was then implemented and was found to perform better than a purely sociohydrological model (SH only model). It was also found to perform better than a purely KPCA based model of yields (KPCA only model). The latter, KPCA only modelling type, is also a standard practice in machine learning approaches to yield predictions at the farm scale. This highlighted that the SH model contributes important sociohydrological processes that cannot be replicated by a purely statistical KPCA model. Leveraging on the results, especially correlation between incomes, yields and model errors, the paper then suggested that approaches such as qualitative interviews with farmers and symbolic regression can be used to infer missing feedbacks and enhance system understanding.

The method relied on farmer surveys to identify KPCA structural error model predictors and to provide much-needed observations of plot scale yields to estimate the residuals. The surveys thus make the KPCA structural error model more local, adding information that the SH only model data lacks because it is driven by remotely sensed or other secondary data sources. Surveys are time consuming and financially costly. Also, the variables that may explain the structural error are not all known *a priori*, and a few iterations are needed to identify missing variables. As such, this method of survey-based structural error modelling needs to be iterative, demanding longitudinal surveys and even more resource commitment. Translating relationships identified in KPCA error models to causal loops, e.g., based on symbolic regression, in the SH model also remains to be explored further. Doing this will enhance the state of the art of SH modelling, where new relationships learned from structural error modeling help unravel novel feedback mechanisms but remain a barrier that is yet to be scaled. Contrasting such an approach could be participatory Causal Loop Diagram (CLD) building, where focus group discussions (FGDs) with farmers are conducted to identify feedback loops that they see as important in determining, e.g., how their farming decisions are affected by water availability and *vice versa*. While such a method will still need to be supplemented with physical water balance models, it can offer an interesting approach that can either be: 1) an alternative to KPCA structural error modelling to identify missing feedbacks, 2) complementary to the KPCA method in identifying novel feedbacks, or 3) supplementary in validating novel feedbacks identified by the KPCA method. However this may also need time-consuming FGD sessions and qualitative analysis and can demonstrate location-to-location variability in assessed feedbacks. Further, it may also be difficult to quantitatively validate the identified feedback loops. Nonetheless, using a mixed method approach to identify missing feedbacks using surveys, KPCA structural error modelling and participatory CLD building offers a way forward in improving SH models.

The methodological approach that was presented here is replicable in the context of human agricultural systems and in other water human systems. The sociohydrological model and the survey can be implemented for any cropping system and location. The sociohydrological model can be implemented with ease, requiring inputs often available from secondary sources. However, household surveys are primary data sources that need human and capital resources. Though the design of the survey presented here followed a standard protocol (Adla, 2023) that can be replicated in different locations, resource constraints may limit the scalability of such an approach. The primary data collection adds value and improves sociohydrological prediction, but at a cost. However, this cost is relatively small compared to the benefits of improved prediction. This predictive value of surveys, which is often significant, justifies such a mixed methods sociohydrological approach.

CRediT authorship contribution statement

Dennis Djohan: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Julien Malard-Adam:** Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Soham Adla:** Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Saket Pande:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships that may be considered as potential competing interests:

Saket Pande reports that financial support was provided by Netherlands Enterprise Agency. Given his role as Associate Editor of Journal of Hydrology, Saket Pande had no involvement in the peer review of this article and had no access to information regarding its peer review. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Thanks are due to Linnaea Cahill for providing constructive comments on the narrative and to Solidaridad Asia Network Limited (SNAL) for supporting farmer surveys.

Appendix A. Supplementary data

Supplementary data for this article can be found online at doi:10.1016/j.jhydrol.2025.134748.

Data availability

All anonymized data and software used in this research are available online on GitHub at https://github.com/julienmalard/tudelf_kpca.

References

Adla, S., Jan 2023. Guidelines to Conduct RANAS Based Socio-Hydrological (SH) Surveys to Understand Behaviour V1.

Adla, S., Aravindakshan, A., Tyagi, A., Guntha, R., Ponce-Pacheco, M.A., Nagi, A., Pastore, P., Pande, S., 2025. Participatory development of mobile agricultural advisory driven by behavioural determinants of adoption. J. Environ.

Manag. 374, 124140. https://doi.org/10.1016/j.jenvman.2025.124140. https://www.sciencedirect.com/science/article/pii/S0301479725001161.

Adla, S., Pande, S., Vico, G., Vora, S., Alam, M.F., Basel, B., Haeffner, M., Sivapalan, M., 2023. Place for sociohydrology in sustainable and climate-resilient agriculture: review and ways forward. Camb. Prisms: Water 1, e13.

Alam, M.F., 2024. Planning Sustainable and Equitable Agricultural Water Management Interventions: an Agent Based Sociohydrology Approach (Dissertation (TU Delft), Delft University of Technology. https://doi.org/10.4233/uuid:cfe2ae4b-0dd0-4aba-a7a4-eed94b20c7d3

Alam, M.F., McClain, M.E., Sikka, A., Daniel, D., Pande, S., 2022. Benefits, equity, and sustainability of community rainwater harvesting structures: an assessment based on farm scale social survey. Front. Environ. Sci. 10, 1043896.

Allen, R.G., Pereira, L.S., Raes, D., 1998. Crop Evapotranspiration-Guidelines for Computing Crop Water Requirements-FAO Irrigation and Drainage Paper 56 Table of Contents (Tech. Rep.). FAO - Food and Agriculture Organization of the United Nations.

Amirkhani, M., Zarei, H., Radmanesh, F., Pande, S., 2022. An operational sociohydrological model to understand the feedbacks between community sensitivity and environmental flows for an endorheic lake basin, Lake Bakhtegan, Iran. J. Hydrol. 605, 127375. https://doi.org/10.1016/j.jhydrol.2021.127375. https://www.sciencedirect.com/science/article/pii/S0022169421014256.

Ayre, M., Collum, V.M., Waters, W., Samson, P., Curro, A., Nettle, R., et al., 2019. Supporting and practising digital innovation with advisers in smart farming. NJAS: Wagening. J. Life Sci. 90-91 (1), 1–12. https://doi.org/10.1016/j.njas.2019.05.001

Beven, K., Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrol. Process. 6 (3), 279–298. https://doi.org/10.1002/hyp.3360060305

Beven, K., Smith, P., Westerberg, I., Freer, J., 2012. Comment on pursuing the method of multiple working hypotheses for hydrological modeling by P. Clark et al. Water Resour. Res. 48 (11), https://doi.org/10.1029/2012WR012282

De, S., Tantri, P., 2014. Borrowing culture and debt relief: evidence from a policy experiment. In: Asian Finance Association (AsianFA) 2014 Conference Paper.

Di Baldassarre, G., Sivapalan, M., Rusca, M., Cudennec, C., Garcia, M., Kreibich, H., et al., 2019. Sociohydrology: scientific challenges in addressing the sustainable development goals. Water Resour. Res. 55 (8), 6327–6355. https://doi.org/10.1029/2018WR023901

Di Baldassarre, G., Viglione, A., Carr, G., Kuil, L., Yan, K., Brandimarte, L., Blöschl, G., 2015. Debates—perspectives on socio-hydrology: capturing feedbacks between physical and social processes. Water Resour. Res. 51 (6), 4770–4781. https://doi.org/10.1002/2014WR016416

Dile, Y., Srinivasan, R., George, C., 2019. QGIS Interface for SWAT (QSWAT).

Fan, S., Brzeska, J., Keyzer, M., Halsema, A., 2013. From Subsistence to Profit: Transforming Smallholder Farms, vol. 26. Intl Food Policy Res Inst, https://doi.org/10.2499/9780896295582

FAO, 2012. AquaCrop: Reference Manual (Tech. Rep. No. June). FAO, Land and Water Division.

Ghoreishi, M., Sheikholeslami, R., Elshorbagy, A., Razavi, S., Belcher, K., 2021. Peering into agricultural rebound phenomenon using a global sensitivity analysis approach. J. Hydrol. 602, 126739. https://doi.org/10.1016/j.jhydrol.2021.126739. https://www.sciencedirect.com/science/article/pii/S0022169421007897.

Giné, X., Kanz, M., 2014. The economic effects of a borrower bailout evidence from an emerging market. Rev. Financ. Stud. 31 (5).

Gu, J., Sun, S., Wang, Y., Li, X., Yin, Y., Sun, J., Qi, X., 2021. Sociohydrology: an effective way to reveal the coupled evolution of human and water systems. Water Resour. Manag. 1–16.

Haeffner, M., 2022. Discussion of “guiding principles for hydrologists conducting interdisciplinary research and fieldwork with participants”. Hydrol. Sci. J. 67 (7), 1145–1148. https://doi.org/10.1080/02626667.2022.2060109

Hair, J.F., 2009. Multivariate Data Analysis.

Hargreaves, G.H., Samani, Z.A., 1985. Reference crop evapotranspiration from temperature. Appl. Eng. Agric. 1 (2), 96–99. https://doi.org/10.13031/2013.26773. http://elibrary.asabe.org/abstract.asp??JID = 3&AID = 26773&CID = aeaj1985&v = 1&i = 2&T = 1.

Hatch, N., Daniel, D., Pande, S., 2022. Behavioral and socio-economic factors controlling irrigation adoption in Maharashtra, India. Hydrol. Sci. J. 67 (6), 847–857. https://doi.org/10.1080/02626667.2022.2058877

Honti, M., Stamm, C., Reichert, P., 2013. Integrated uncertainty assessment of discharge predictions with a statistical error model. Water Resour. Res. 49 (8), 4866–4884. https://doi.org/10.1002/wrcr.20374

Hussain, S., Ahmad, A., Wajid, A., Khaliq, T., Hussain, N., Mubeen, M., et al., 2020. Irrigation scheduling for cotton cultivation. Cotton Prod. Uses: Agron., Crop Prot., Postharvest Technol. 59–80.

Jabed, M.A., Azmi Murad, M.A., 2024. Crop yield prediction in agriculture: a comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. Heliyon 10 (24), e40836. https://doi.org/10.1016/j.heliyon.2024.e40836. https://www.sciencedirect.com/science/article/pii/S2405844024168673.

Janssen, J.M., 2020. Estimating New Reservoir Locations with the Use of a Hydrological Model for Small Holder Cotton Farmers in Maharashtra, India. Estimating New Reservoir Locations with the Use of a Hydrological Model for Small Holder Farmers in Maharashtra, India (Tech. Rep.). TU Delft.

Joshi, P.K., Tyagi, N.K., 2019. Small farm holders and climate change: overcoming the impacts in India. In: Climate Smart Agriculture in South Asia: Technologies, Policies and Institutions. Springer, pp. 49–72.

Kamble, R.K., Tikadar, K.S., 2024. Cotton cultivating marginalised farmers’ climate change perceptions, impacts and adaptation strategies in Vidarbha region, central India. Sustain. Agri Food Environ. Res. 12 (1), https://portalrevistas.uct.cl/index.php/safer/article/view/2430.

Kanz, M., 2012. What does debt relief do for development? evidence from India ’ s bailout program for highly-indebted rural households. Am. Econ. J.: Appl. Econ. 8 (4).

Katti, R.K., 1979. Search for solutions to problems in black cotton soils. Indian Geotech. J. 9 (1), 1–82.

Kennedy, M.C., O’Hagan, A., 2001. Bayesian calibration of computer models. Stat. Methodol. 63 (3), 425–464.

Kerala Legislative Assembly, 2012. The kerala farmers’ debt relief commission (amendment) bill, 2012. Thiruvananthapuram, Thirteenth Kerala Legislative Assembly, http://www.egazette.kerala.gov.in/pdf/2012/15/Part_2/farmers.pdf.

Khan, A., Ansari, S.A., 2023. Efficiency in cotton production across the states in India. Bhartiya Krishi Anusandhan Patr. 38 (2), 138–144.

Kreibich, H., Sivapalan, M., AghaKouchak, A., Addor, N., Aksoy, H., Arheimer, B., et al., 2025. Panta Rhei: a decade of progress in research on change in Hydrology and society. Hydrol. Sci. J. 70 (8), 1–27. https://doi.org/10.1080/02626667.2025.2469762

Li, B., Zheng, Y., Di Baldassarre, G., Xu, P., Pande, S., Sivapalan, M., 2023. Groundwater vulnerability in a megacity under climate and economic changes: a coupled sociohydrological analysis. Water Resour. Res. 59 (12), e2022WR033943. https://doi.org/10.1029/2022WR033943

Mishra, S., 2006. Suicide of Farmers in Maharashtra (Tech. Rep.). Indira Gandhi Institute of Development Research.

Munoz, J.M., Udrescu, S.M., Garcia Ruiz, R.F., Mar 2025. Discovering nuclear models from symbolic machine learning. Commun. Phys. 8 (1).

Nash, J.E., Sutcliffe, J.V., Apr 1970. River flow forecasting through conceptual models part I—a discussion of principles. J. Hydrol. 10 (3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Pai, D.S., Sridhar, L., Rajeevan, M., Sreejith, O.P., Satbhai, N.S., Mukhopadyay, B., 2014. Development of a new high spatial resolution (0.25° × 0.25°) long period (1901–2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region data sets of different spatial resolutions and time period. MAUSAM 1 (1), 1–18.

Pande, S., Savenije, H.H.G., Mar 2016. A sociohydrological model for smallholder farmers in Maharashtra, India. Water Resour. Res. 52 (3), 1923–1947. https://doi.org/10.1002/2015WR017841

Pande, S., Sivapalan, M., 2017. Progress in socio-hydrology: a meta-analysis of challenges and opportunities. Wiley Interdiscip. Rev.: Water 4 (4), e1193. https://doi.org/10.1002/wat2.1193

Pathak, A., Chattopadhyay, A.K., 2021. Debates on agricultural loan waiver schemes in India: myths and realities. Contemp. South Asia 29 (4), 560–570.

Pedregosa, F., Weiss, R., Brucher, M., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pianosi, F., Raso, L., 2012. Dynamic modeling of predictive uncertainty by regression on absolute errors. Water Resour. Res. 48 (3), https://doi.org/10.1029/2011WR010603

Ponce-Pacheco, M.A., Adla, S., Guntha, R., Aravindakshan, A., Presannakumar, M., Tyagi, A., et al., 2025. Makara: a tool for cotton farmers to evaluate risk to income. Smart Agric. Technol. 10, 100759. https://doi.org/10.1016/j.atech.2024.100759. https://www.sciencedirect.com/science/article/pii/S2772375524003630.

QGIS Development Team, 2021. QGIS geographic information system. https://www.qgis.org.

Quesnel, K.J., Ajami, N.K., 2017. Changes in water consumption linked to heavy news media coverage of extreme climatic events. Sci. Adv. 3 (10), e1700784. https://doi.org/10.1126/sciadv.1700784

Rachunok, B., Fletcher, S., 2023. Socio-hydrological drought impacts on urban water affordability. Nature Water 1 (1), 83–94.

Raghavendra, T., Reddy, Y.R., 2020. Physiological determinants and yield components as influenced by high density planting system in cotton. Int. J. Curr. Microbiol. App. Sci. 9 (4), 748–754.

Raschka, S., Mirjalili, V., 2019. Python Machine Learning, third ed. Packt Publishing.

Ratna, S.B., 2012. Summer monsoon rainfall variability over Maharashtra, India. Pure Appl. Geophys. 169 (1–2), 259–273. https://doi.org/10.1007/s00024-011-0276-4

RBI, 2009. Agricultural debt waiver and debt relief scheme, 2008. Reserve Bank of India. https://rbi.org.in/scripts/BS_CircularIndexDisplay.aspx?Id = 4190.

Roobavannan, M., Kandasamy, J., Pande, S., Vigneswaran, S., Sivapalan, M., Oct 2017. Role of sectoral transformation in the evolution of water management norms in agricultural catchments: a sociohydrologic modeling analysis. Water Resour. Res. 53 (10), 8344–8365. https://doi.org/10.1002/2017WR020671

Samani, Z., Jul 2000. Estimating solar radiation and evapotranspiration using minimum climatological data. J. Irrig. Drain. Eng. 126 (4), 265–267. https://doi.org/10.1061/(asce)0733-9437(2000)126:4(265). https://ascelibrary.org/doi/abs/10.1061/%28ASCE%290733-9437%282000%29126%3A4%28265%29.

Schölkopf, B., Smola, A., Müller, K.-R., 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. 10 (5).

Shwetha, M.N., Shakuntala, I., Lavanya, T., Suhasini, K., Meena, A., 2022. Growth in area, production and productivity of cotton crop in India: a state-wise analysis. Int. J. Environ. Clim. Change 12 (11), 51–57.

Sivapalan, M., Blöschl, G., 2015. Time scale interactions and the coevolution of humans and water. Water Resour. Res. 51 (9), 6988–7022. https://doi.org/10.1002/2015WR017896

Sravanth, K.R.S., Sundaram, N., 2019. Agricultural crisis and farmers suicides in India. Int. J. Innov. Technol. Explor. Eng. 8 (11), https://doi.org/10.35940/ijitee.K1855.0981119

Srinivasan, V., Sanderson, M., Garcia, M., Konar, M., Blöschl, G., Sivapalan, M., 2018. Moving socio-hydrologic modelling forward: unpacking hidden assumptions, values and model structure by engaging with stakeholders: reply to "what is the role of the model in socio-hydrology?". Hydrol. Sci. J. 63 (9), 1444–1446. https://doi.org/10.1080/02626667.2018.1499026

Srivastava, A.K., Rajeevan, M., Kshirsagar, S.R., 2009. Development of a high resolution daily gridded temperature data set (1969–2005) for the Indian region. Atmos. Sci. Lett. 10 (4), 249–254. https://doi.org/10.1002/asl

van Wirdum, C., Hatch, N., Mohammed Yasir Abbas Mohammed Ali, M., Raghunathan, P., Willard, T., 2019. Multidisciplinary project Cotton Water: baseline study of designing sustainable instruments for smallholders in Maharashtra, India. http://resolver.tudelft.nl/uuid:16fc0b0b-72e6-47da-9a91-2305adf65e58.

Xu, T., Valocchi, A.J., 2015. A Bayesian approach to improved calibration and prediction of groundwater models with structural error. Water Resour. Res. 51 (11), 9290–9311. https://doi.org/10.1002/2015WR017912