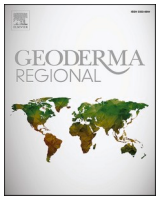
Geoderma Regional 37 (2024) e00784

Contents lists available at ScienceDirect

Geoderma Regional

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



Characterizing soil hydrology in the Indo-Gangetic plain of Bihar, India: Methods and preliminary results

D.G. Rossiter [a,*], Laura Arenas-Calle [a], Anton Urfels [c], Harishankar Nayak [a], Sonam Sherpa [b], Andrew McDonald [a]

[a] Section of Soil & Crop Sciences, School of Integrative Plant Sciences, Bradfield Hall, Cornell University, Ithaca, NY 14853, USA
[b] CIMMYT CSISA, Patna, Bihar 800025, India
[c] International Rice Research Institute, Los Baños, Philippines

ARTICLE INFO

Keywords:
Hydrologic soil classification
Landscape diagnostic survey
WRB Cambisols, Fluvisols, Vertisols

ABSTRACT

In the Eastern Gangetic Plain (EGP) soil hydrology is a major determinant of land use and also governs the ecosystem services derived from cropping systems, particularly greenhouse gas (GHG) emissions from rice fields. To characterize patterns of soil hydrology in these, daily field monitoring of water levels was conducted during the monsoon (*kharif*) season in a comparatively wet (2021) and dry (2022) year with flooding depth and drainage tracked with field water tubes across 47 (2021) and 183 (2022) locations. Fields were clustered into hydrologic response types (HRT) which can then be used for land surface modelling, land use recommendations, and to target agronomic interventions that contribute to sustainable development outcomes. Clusters based on two methods of summarizing a single information source were compared. The information source was a time-series of field water-level observations, and the two methods were (1) the original time-series and their first differences and (2) a set of derived hydrologic descriptors that are conceptually related to greenhouse gas (GHG) emissions. Clustering was (1) by k-means with an optimization of cluster numbers and (2) by hierarchical clustering with the same number of clusters as identified by k-means. Hydrologic behaviour shifted dramatically between growing seasons, and it was not possible to identify consistent HRT's across years. The clusters had only a weak relation with soil properties, almost no relation with farmer perception of relative landscape position, and no relation with rice establishment method. Clusters based on time-series were moderately well predicted in the dry year 2022 by optimized random forest models, with the most important predictors being the number of irrigations, seasonal precipitation, pre-monsoon groundwater levels, seasonal groundwater level change, and pH, this latter as a surrogate for landscape position and other soil properties. In the wet year 2021 clusters were (poorly) predicted by just seasonal precipitation and pre-monsoon groundwater levels. This shows the complex relation of soil hydrology with landscape position and land management, as well as synoptic climate. By contrast, clusters based on the descriptors were not well-matched with those from the time-series, and could not be well predicted by random forest models. This shows that different clustering criteria may result in different interpretations of the landscape hydrology and thus different heuristics for anticipating the hydrology of a given field under different management choices.

In the Indo-Gangetic plain (IGP) soil hydrology is a major determinant of the spatial distribution of land use systems. These are mainly rice-based systems: dominantly rice-wheat with rice-pulse and rice-maize annual rotations also practised in the monsoon (*kharif*) and dry winter seasons (*rabi*), respectively. In the lower Ganges Plain, double rice crops are also common (Timsina and Connor, 2001). Rice is especially adapted to saturated soil conditions and periodic shallow flooding that other crops without physiological traits like aerenchyma cannot survive (Miro and Ismail, 2013). As such, soil hydrology during the monsoon season governs the distribution and production risks faced by rice and non-rice crops (Urfels et al., 2021; McDonald et al., 2006). The source of soil water (rain, floods, lateral flow, through flow), inter-annual variation, seasonal pattern, and within-season time series of water levels in and above the soil greatly influence crop management (e.

* Corresponding author.
E-mail addresses: d.g.rossiter@cornell.edu (D.G. Rossiter), la397@cornell.edu (L. Arenas-Calle), a.urfels@irri.org (A. Urfels), hsn28@cornell.edu (H. Nayak), s.sherpa@cgiar.org (S. Sherpa), andrew.mcdonald@cornell.edu (A. McDonald).

https://doi.org/10.1016/j.geodrs.2024.e00784
Received 8 December 2023; Received in revised form 6 March 2024; Accepted 8 March 2024
Available online 16 March 2024

Table 1
Soil properties, 2021 (left), 2022 (right); sand, silt, clay, OC: g/dg; EC: ds/m.

	sand	silt	clay	pH	EC	OC	sand	silt	clay	pH	EC	OC
Min.	2	0	0	5.8	0.00	0.19	12	10	5	5.3	0.10	0.16
st Qu.	12	3	14	8.0	0.28	0.45	32	28	15	6.7	0.15	0.41
Median	18	6	56	8.2	0.34	0.57	40	32	25	7.4	0.19	0.55
Mean	21	28	50	8.2	0.45	0.57	41	34	25	7.4	0.21	0.52
rd Qu.	28	58	78	8.5	0.45	0.64	50	39	32	8.1	0.24	0.62
Max	66	87	94	8.8	2.45	0.91	72	68	60	9.0	0.83	0.87

Table 2
Summary of descriptors, 2021. See text for descriptor codes and names.

	min	max	range	IQR	Q0.25	Q0.75
days.flood.pct	9.09	100.00	90.91	44.55	33.64	78.18
flood.0.5.cm	0.00	48.18	48.18	16.82	6.36	23.18
flood.5.10.cm	0.00	33.64	33.64	8.18	1.82	10.00
flood.10.cm	0.00	100.00	100.00	62.27	1.82	64.09
flood.events	1.00	9.00	8.00	4.50	2.00	6.50
flood.events.one.day	0.00	71.43	71.43	25.00	0.00	25.00
flood.events.one.week	0.00	100.00	100.00	53.57	0.00	53.57
flood.events.one.month	0.00	66.67	66.67	36.67	0.00	36.67
flood.events.more.month	0.00	100.00	100.00	50.00	0.00	50.00
flood.5.0.cm	0.00	37.27	37.27	10.91	2.73	13.64
flood.10.5.cm	0.00	10.00	10.00	5.00	0.91	5.91
flood.10.minus.cm	0.00	73.64	73.64	30.00	14.55	44.55
avg.duration.dry.10.minus	0.00	34.00	34.00	11.88	2.50	14.38
avg.duration.dry.10.5	0.00	3.50	3.50	0.45	1.00	1.45
avg.duration.dry.5.0	0.00	16.67	16.67	3.29	2.96	6.25
days.redox	0.00	6.00	6.00	2.50	0.00	2.50
days.quick.drain	0.00	2.00	2.00	0.00	0.00	0.00

Table 3
Summary of descriptors, 2022. See text for descriptor codes and names.

	min	max	range	IQR	Q0.25	Q0.75
days.flood.pct	0.00	70.83	70.83	31.88	10.00	41.88
flood.0.5.cm	0.00	43.33	43.33	18.54	5.00	23.54
flood.5.10.cm	0.00	36.67	36.67	11.04	0.83	11.88
flood.10.cm	0.00	50.00	50.00	5.83	0.00	5.83
flood.events	0.00	27.00	27.00	6.00	3.00	9.00
flood.events.one.day	0.00	100.00	100.00	49.52	8.33	57.86
flood.events.one.week	0.00	100.00	100.00	60.00	0.00	60.00
flood.events.one.month	0.00	100.00	100.00	22.92	0.00	22.92
flood.events.more.month	0.00	100.00	100.00	0.00	0.00	0.00
flood.5.0.cm	0.83	31.67	30.83	11.67	4.17	15.83
flood.10.5 cm	0.00	22.50	22.50	8.54	3.96	12.50
flood.10.minus.cm	17.50	98.33	80.83	35.00	36.67	71.67
avg.duration.dry.10.minus	0.00	45.50	45.50	6.05	3.95	10.00
avg.duration.dry.10.5	0.00	8.00	8.00	0.83	1.17	2.00
avg.duration.dry.5.0	1.00	5.50	4.50	1.05	1.16	2.21
days.redox	0.00	26.00	26.00	4.00	2.00	6.00
days.quick.drain	0.00	26.00	26.00	1.00	0.00	1.00

g. timing of crop establishment) as well as yield outcomes (McDonald et al., 2022) Soil hydrology is also a major determinant of greenhouse gas (GHG) emissions from these systems (Kraus et al., 2015). Many global assessments of GHG emissions in rice assume that systems are flooded for the duration of the cropping season, and that major reductions in GHG emissions are achievable if water management is changed to favor periodic or occasional drainage (Bo et al., 2022). Accurate modelling of soil water dynamics is crucial to simulate GHG-drivers parameters such as pH, redox potential (Eh), and substrate concentrations (Yin et al., 2020).

The landscape of the IGP has clear differences in soil hydrology due to local landscape position and the relation to fluvial systems (Sinha et al., 2005) and groundwater resources (Bonsor et al., 2017). This is recognized by the local perception of landscape position as “upland”, “medium land”, “lowland”, although these terms have no precise definition and vary across sub-regions. Soil profile characteristics (e.g., hydraulic conductivity, porosity, bulk density, layering) also affect soil hydrology, irrespective of landscape position. Some of these effects can be seasonal or transient in nature. For example, temporal differences in hydraulic conductivity may control early-season patterns of soil moisture variation but have no influence on the field water balance later in the season, when landscape and shallow groundwater factors exert more influence. Management practices including irrigation, puddling and compaction also have impacts on dynamic soil properties that can vary within a season, notably water balance and hydraulic conductivity varying with degree of saturation (McDonald et al., 2006). The totality of the time series of soil hydrology is termed the *soil hydrological response*. This is expressed both within and between seasons. Clusters of sites with similar soil hydrology are termed *hydrological response types* (HRT).

The hydrology of the Bihar portion of the eastern IGP is complex. The Ganges River flows across the centre of the State, and has several major tributaries coming from the Himalayan foothills to the north and Jharkand State to the south (Jain and Sinha, 2003). These have formed terraces and abandoned channels, and are further drained by smaller streams and canals. These factors contribute to the observed spatial variability of the time-series of soil-water relations.

The objective of this work is to identify and characterize soil hydrological response types (HRT) for the cropland of the IGP in Bihar state, where rice is the *kharif* component. This landscape segmentation is intended for use in models where soil hydrology dynamics are important, e.g., GHG emission models. Once identified, the HRT could be used to select representative sites for modelling, allowing extrapolation of results to other sites with the same HRT. The HRT can also be used for suitability evaluation for alternative rice-based systems, including new technologies (e.g., direct-seeded rice, adjusted planting dates, variety maturity class) or management strategies, including opportunities for irrigation-led intensification (Balwinder-Singh et al., 2019).

This work is part of the Cereal Systems Initiative for South Asia (CSISA) of the International Maize and Wheat Improvement Center (CIMMYT) in partnership with the International Food Policy Research Institute (IFPRI), the International Rice Research Institute (IRRI) and the International Water Management Institute (IWMI) (CIMMYT, 2017). Among CSISA's goals are (1) to accelerate widespread adoption of resource-conserving practices, technologies and services that increase yields with lower water, labour and input costs and (2) to disseminate new knowledge on cropping system management practices that can withstand the impacts of climate variability and change. To achieve these goals, CSISA and the Indian Council of Agricultural Research (ICAR) have initiated a partnership to characterize crop production systems at the national scale (McDonald et al., 2023). This initiative, termed the ‘Landscape Crop Assessment Survey (LCAS)’ endeavours to enrich basic data stacks on yield and agronomic production practices with environmental data, including weather, soils, and hydrologic characterization information.

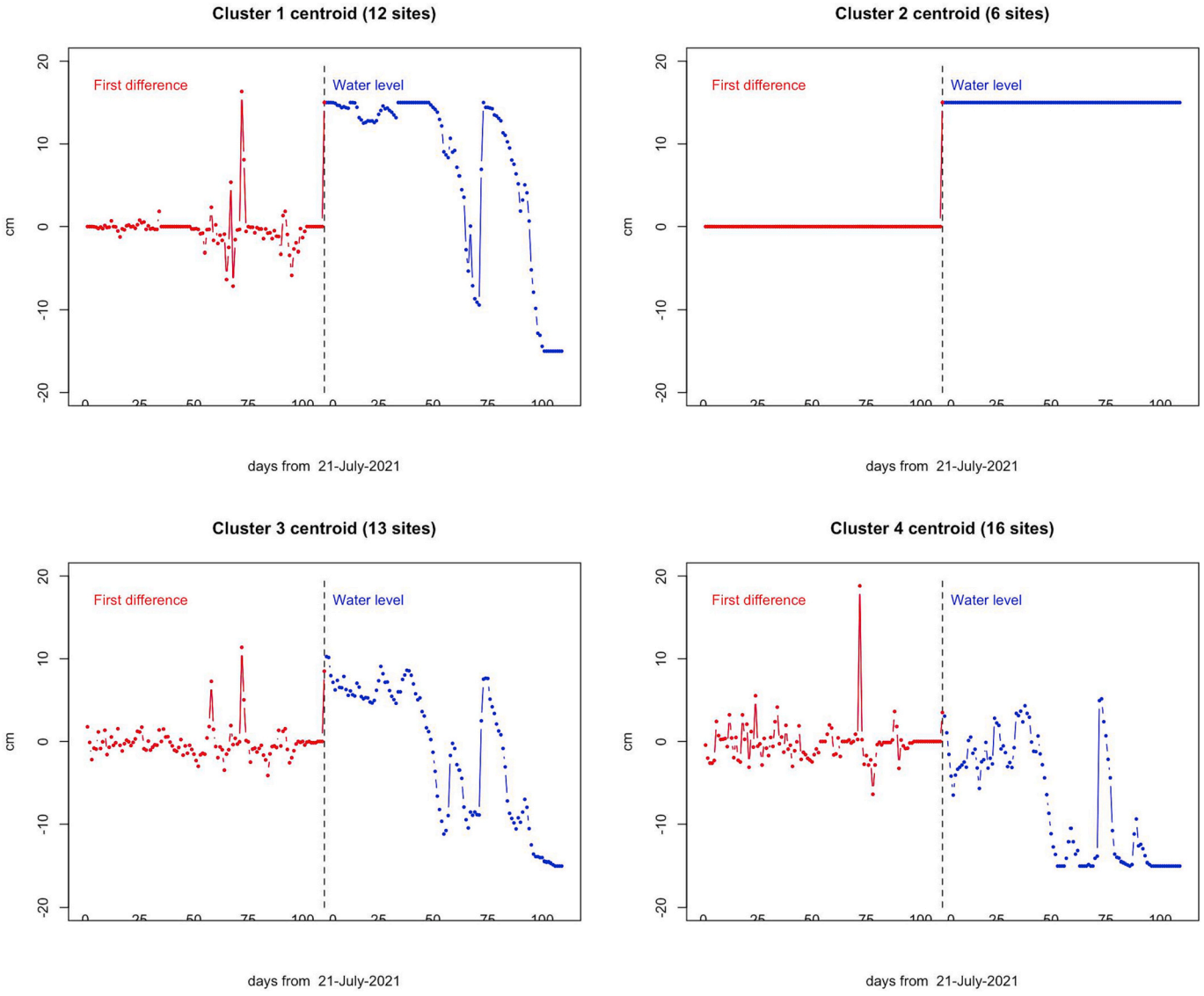


Fig. 1. Cluster centroids 2021, first differences (left half) and water levels (right half).

1. Soil hydrologic response types

Several efforts have been made to group soils by their hydrologic behaviour. An early example is the "hydrology of soil types" (HOST) system developed for the United Kingdom (Boorman et al., 1995; Lilly et al., 1998). HOST groups soils by the dominant pathways and rate of water movement through soil, and was designed primary for watershed hydrology modelling. Another common classification is the USDA's hydrologic soil groups (HSG) (USDA Natural Resources Conservation Service, 2020), which have been mapped at Global scale (Ross et al., 2018) and in smaller regions (Faouzi et al., 2023) in support primarily of rainfall-runoff relations for RUSLE-based soil erosion modelling using the curve number (CN) method (Garen and Moore, 2005) or as an enhancement to DNDC using the related MUSLE soil erosion model (Deng et al., 2011). Classifications for specific model types include those of Quisenberry et al. (1993), who grouped South Carolina (USA) soils based on their behaviour for water and chemical transport. The close relation between soil hydrology and landscape position within a watershed was proven by, among others, Park and van de Giesen (2004). Landscape segmentation of wetlands is well-developed as a hydrogeomorphic (HGM) classification (Semeniuk and Semeniuk, 2018) based on geomorphic setting, dominant water source and transport, and hydrodynamics of moving water.

However, in the IGP, neither watershed nor soil erosion modelling are of primary importance, and most of the agricultural area is not developed from reclaimed wetlands, rather from alluvial plains and terraces. Here the interest is in the time sequence over the year of presence of water in and over the soil profile, potentially influenced by the dynamic interactions between landscape position, soil properties, groundwater, irrigation, and soil management. Hence a new approach is needed.

2. Study area

The Indian state of Bihar is at the eastern end of the IGP, bisected by the Ganges from west to east, and with several major tributaries: the Sone, Ghaghara, Gandak and Koshi. The State is of low relief except for some hilly areas bordering Nepal to the northwest and Jharkhand to the south. Agriculture is the dominant land use, with rice-wheat systems in the wet (*kharif*) and dry (*rabi*) seasons as the dominant land use. Maize and pulses systems are found in some areas in the winter cropping cycle. A wide variety of secondary crops are grown, especially in areas with good access to major markets such as the capital city Patna. This State is a priority area for the activities of CSISA. In this study only the areas

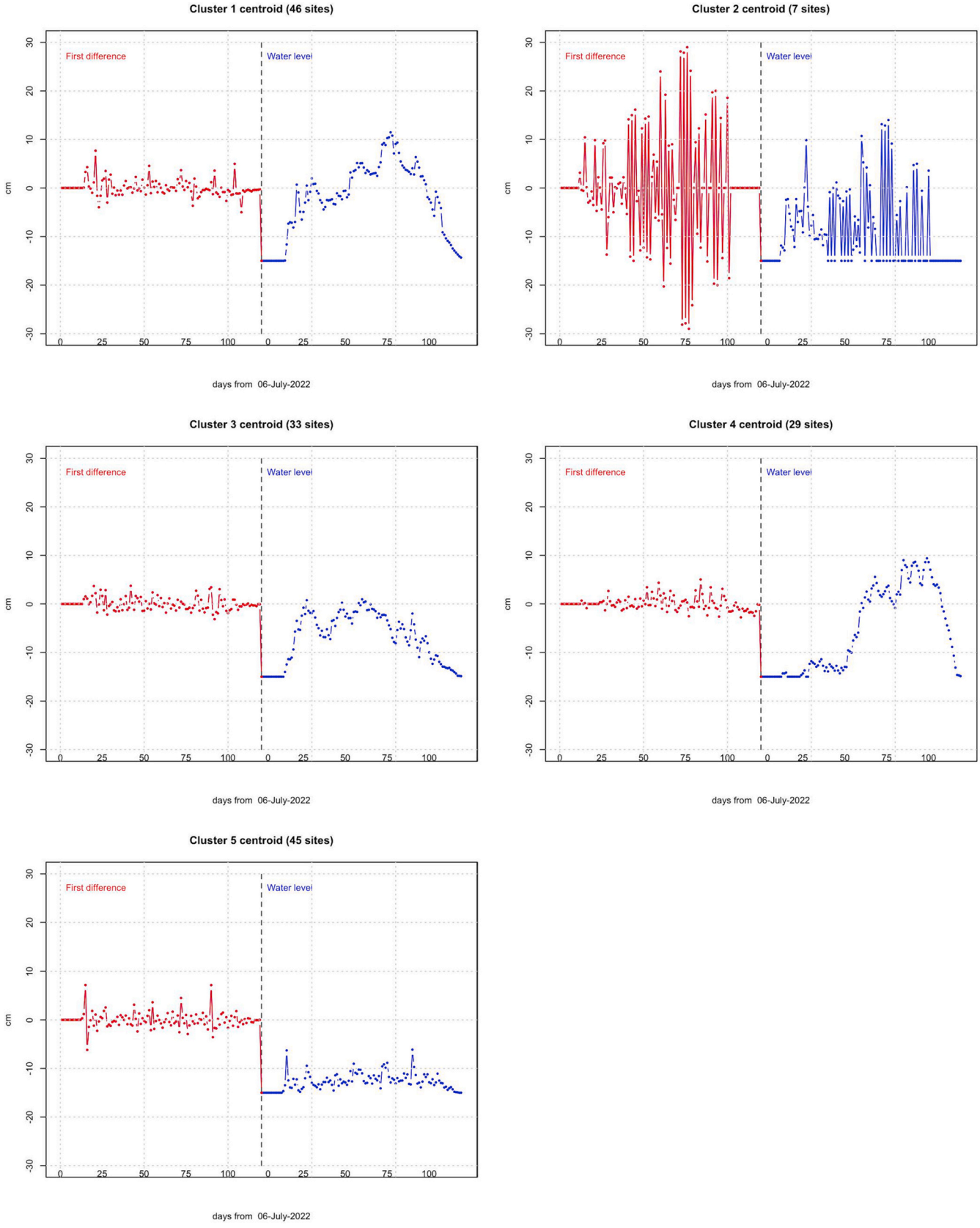


Fig. 2. Cluster centroids 2022, first differences (left half) and water levels (right half).

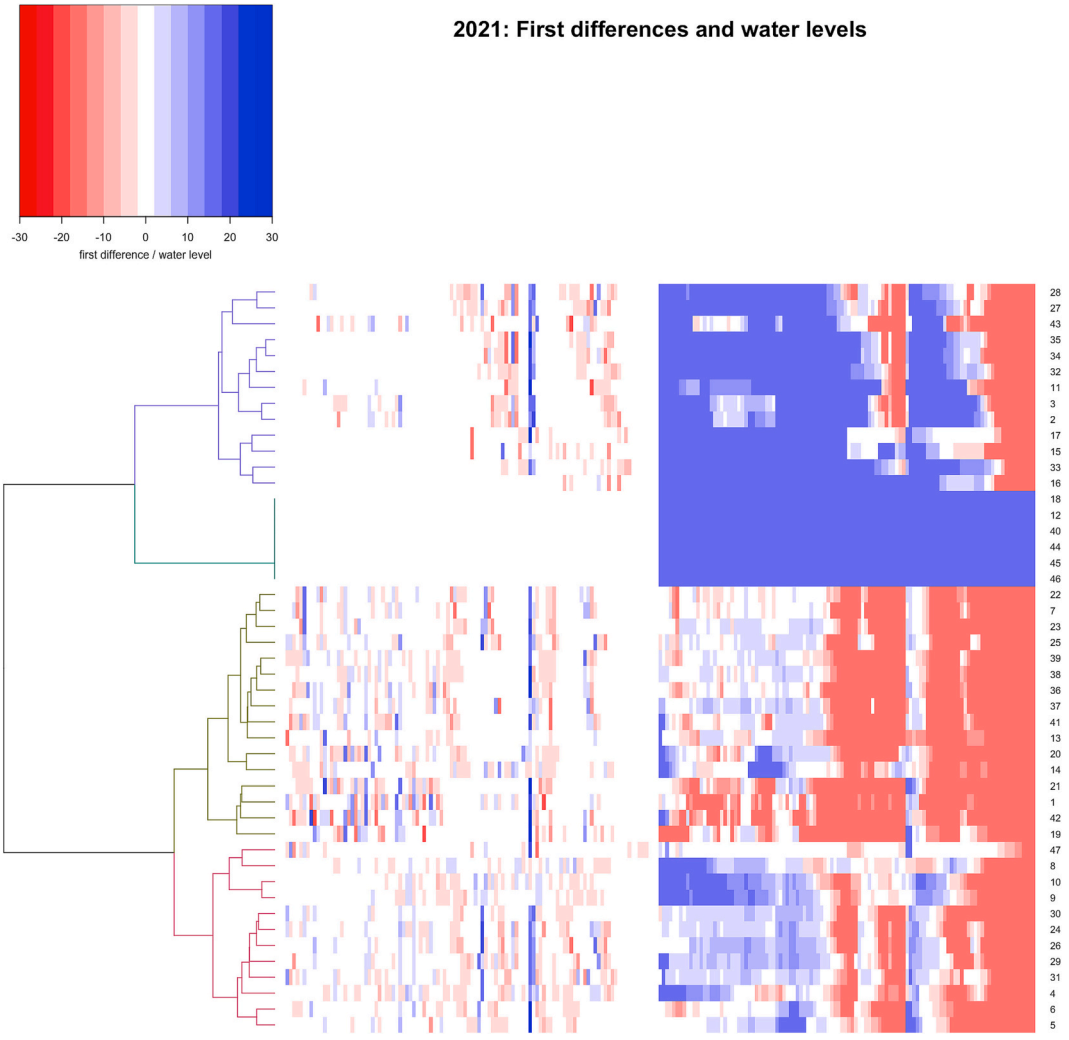


Fig. 3. Hierarchical clustering of time series, 2021.

planted to *kharif* rice are characterized.

3. Data sources

The identification of HRT is based on clustering observations from direct field measurements. Observations from two contrasting *kharif* seasons (July – October) were used to determine if HRT are consistent over years, or must be recalculated for each year. Year 2021 had more precipitation than long-term (1901–2017) area-weighted statistics (904 mm, standard deviation 156 mm) for the four *kharif* months (June–September) in Bihar (Government of India, 2022). In this year the *kharif* rainfall over six stations near the field locations had a mean of 474 mm and a standard deviation of 130 mm for the three months recorded (July–September), however June 2021 had so much rainfall that gauges were not monitored and rice could not be established. By contrast for 2022 the 23 rainfall gauges averaged only 354 mm for the season. This was less than any previous amount in the 1901–2017 period..

Field observations were made as part of the Landscape Crop Assessment Surveys (LCAS). This survey included farmer-reported information on hydrology-related management practices at each site: rice establishment method (2022 only), number of irrigations, timing of irrigations (early vegetative, mid vegetative, flowering and grain filling), reason for irrigating (visible crop stress, soil cracking, disappearance of flood water, crop growth stage), tubewell depth, pump type and power. In 2021 the management practices were consistent among sites, likely due to the very wet conditions than limited options and made irrigation irrelevant.

At 47 (2021) and 160 (2022) LCAS points, a set of so-called *pani-pipes* ("pani" is the Hindi word for "water") were installed in farmer's fields during the *kharif* season in 2021 and 2022. These are 30 cm long by 10 cm diameter plastic tubes inserted to a depth of 15 cm in the soil, so that 15 cm are above the soil surface, with 5 cm diameter holes, spaced 2 cm apart (edge to edge) both horizontally and vertically, in the section below the soil surface.

At each pani-pipe location the farmer was asked for the perceived landscape position, one of "upland", "medium land", "lowland". These terms refer not to the actual local elevation, but rather the farmers' perception of relative differences in hydrology in a local context that also pertains to the types of crops that can be cultivated during the wet kharif season. "Lowland" is often flooded with poor drainage; "upland" allows a full range of crops, with irrigation if necessary and also good drainage; "medium land" is intermediate. These terms are not used consistently across the study area, but are commonly-used local terms.

In 2022 many sites were paired or closely-spaced, for comparison of direct-seeded rice (DSR) and transplanted rice (TPR). In the smaller study of 2021 all sites were TPR. Pipes were installed at representative locations in the field at the technician's discretion. The water levels in the pipes were measured each morning during the cropping season. The measured levels thus ranged between −15 to +15 cm from the soil

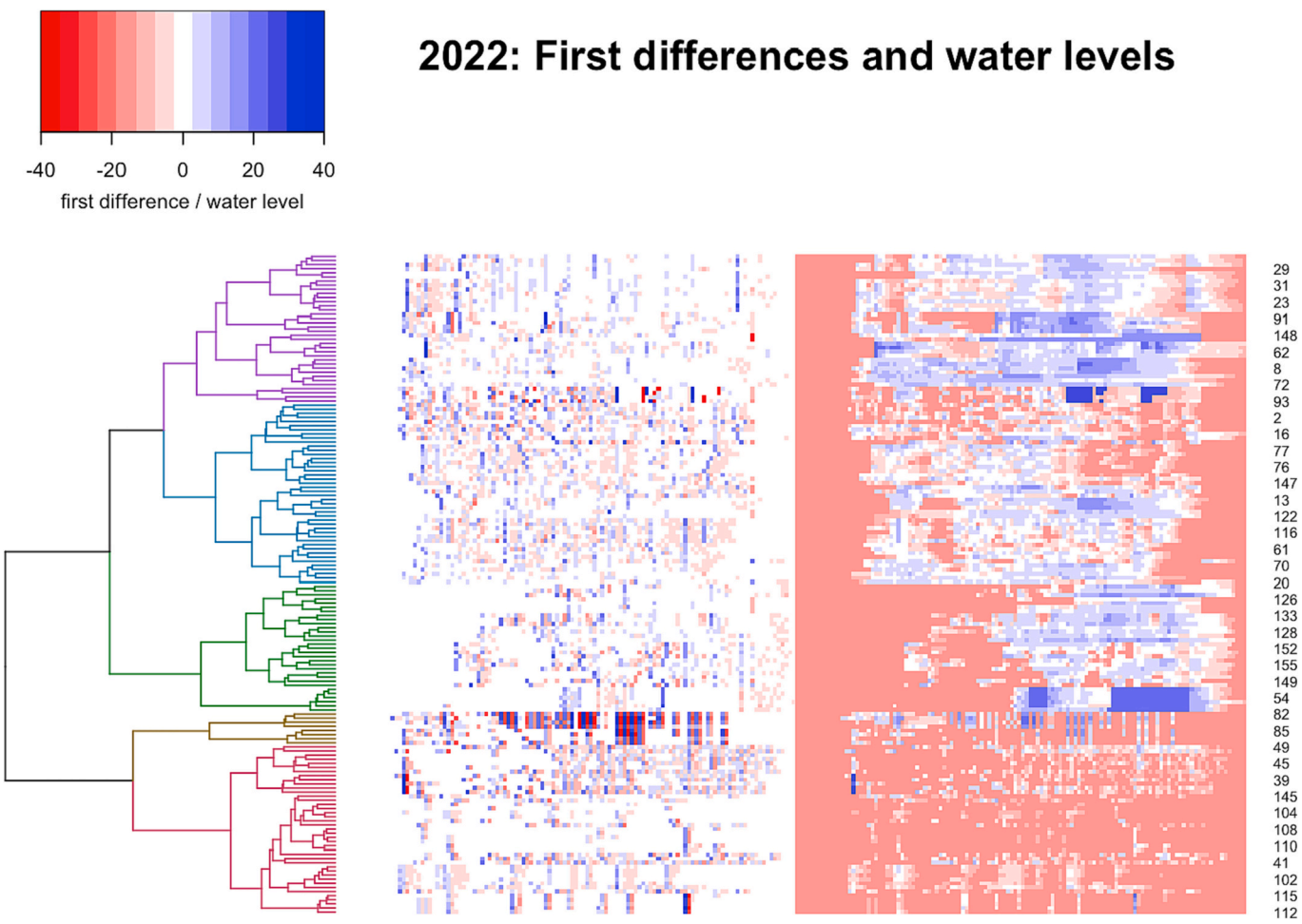


Fig. 4. Hierarchical clustering of time series, 2022.

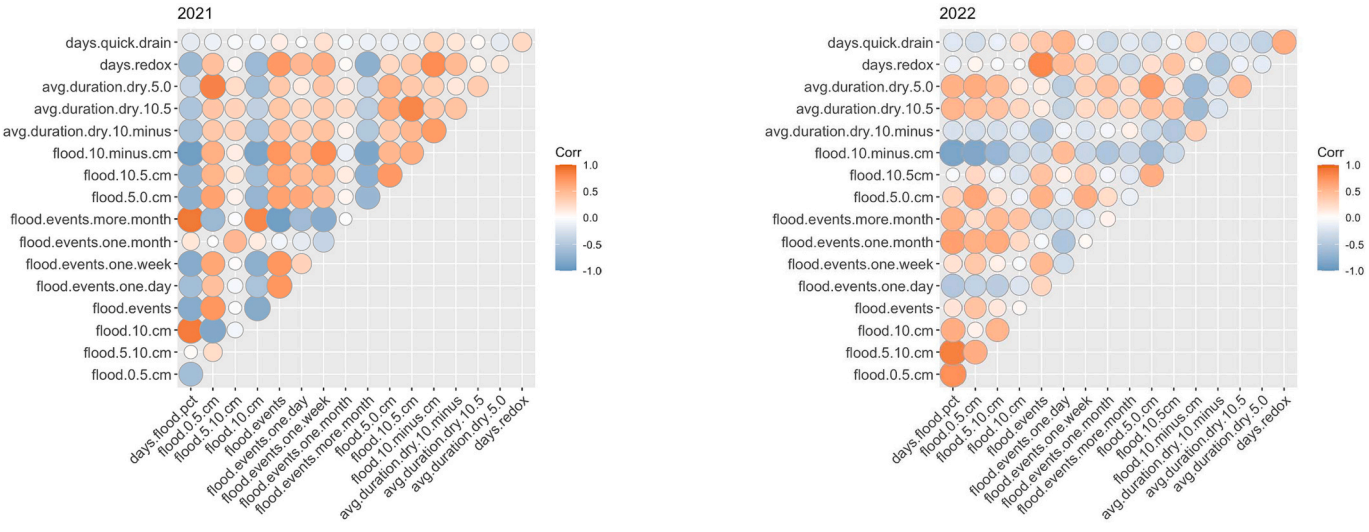


Fig. 5. Spearman's rank correlation of descriptors, 2021 (left), 2022 (right).

surface. In 2022 water levels above the pipe top were also measured. The pani-pipes show the water level, but not the soil moisture status in the 0 to −15 cm range. In addition, rain gauges (2021: 6, 2022: 23) were installed near sets of pani-pipes and read daily.

The 2021 dataset consists of 47 pani-pipes, all in West Champaran district in northwestern Bihar (Fig. 9). These were read from 21-July through 07-November-2021. The 2022 dataset consists of 160 pani-pipes from 15 districts of Bihar and three districts of the adjacent eastern Uttar Pradesh State (Fig. 10). These were read from 06-July until rice harvest in mid-October-2022.

Groundwater levels pre- and post-monsoon (2021: mid-January and late November; 2022 late May and early November) were obtained from

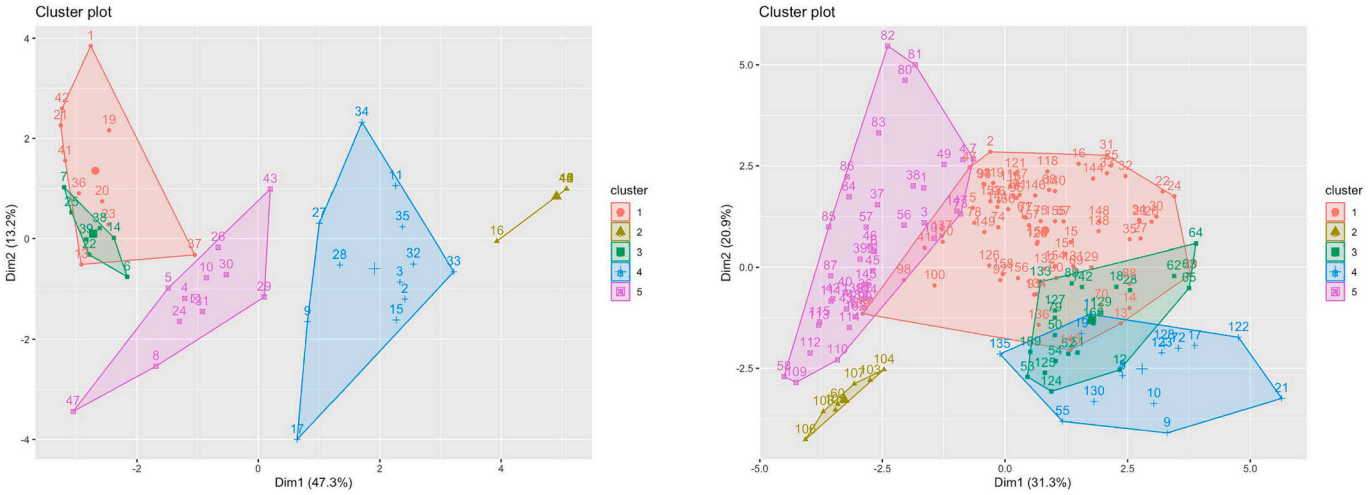


Fig. 6. k-means clusters based on descriptors in principal component space (PC1, PC2), 2021 (left), 2022 (right).

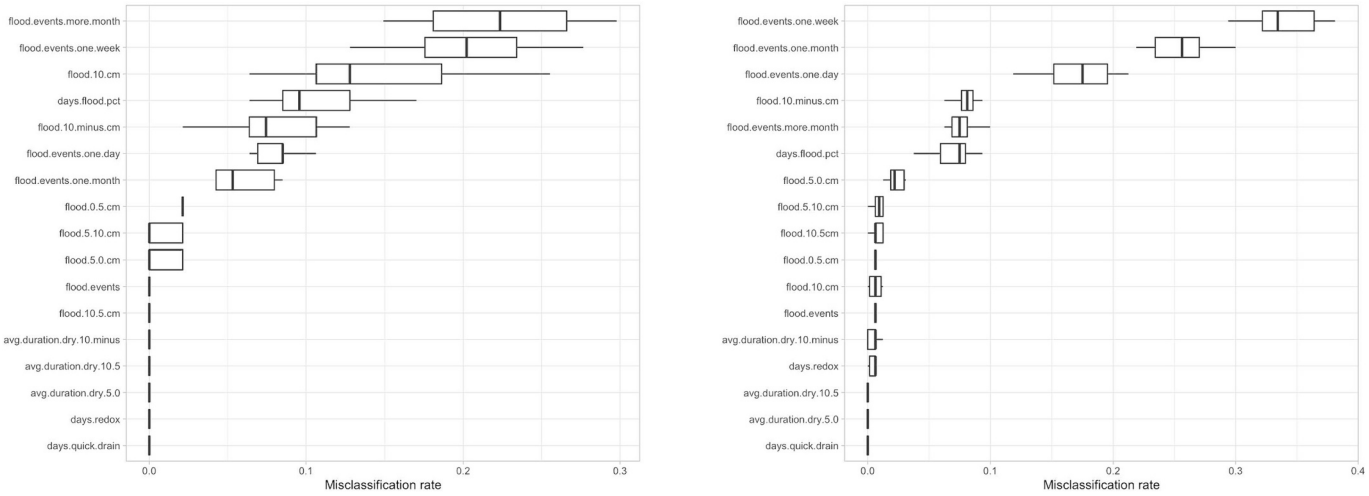


Fig. 7. Feature importance of descriptors for k-means clustering, 2021 (left), 2022 (right). Boxplots show the distribution among trees in the random forest.

the National Water Informatics Centre (NWIC) for 596 (2021) and 781 (2022) monitoring wells distributed over Bihar and eastern UP. Each pani-pipe location was matched with its closest well using the st_distance function of the sf R package. This distance ranged from 2400 to 24,000 m, median 11,450 m (2021, 10 wells) and 650 to 16,300 m, median 4900 m (2022, 34 wells). Using the same method, each pani-pipe was matched with its closest rain gauge, for which the in-season total rainfall was computed. This distance ranged from 46 to 45,600 m, median 4040 m (2021, 6 gauges) and 0 to 35,600 m, median 810 m (2022, 23 gauges).

The terrain elevation of each pani-pipe was extracted from the Amazon Web Services Terrain Tile service (Amazon Web Services, 2023) using the elevatr R package, at zoom level 11 (approximately 120 m horizontal resolution). Elevation differences were not large in this plain area, but did range from 63 to 120 m.a.s.l. (2021) and 38–117 m.a.s.l. (2022). Differences within the 120 m tile in these rice-growing areas are minor in comparison, around 3 m maximum.

At the pani-pipe location soil samples were also collected and analysed in the soil chemistry laboratory of the Dr. Rajendra Prasad Central Agricultural University, Pusa, Bihar, and the Bihar Agricultural University, Sabour, Bihar, for the 2021 and 2022 datasets, respectively. Table 1 summarizes the basic soil properties. There was a fairly wide total and inter-quartile range for most properties.

4. Methods

All computation was with packages of the R environment for statistical computing (R Core Team, 2023). An R Markdown (Allaire et al., 2023) document was developed to ensure reproducible results.

4.1. Characterizing time series of pani-pipe measurements

The time series of water levels in the pani-pipes were considered two ways: (1) as a combined time series of water levels and their first differences; (2) as a set of descriptors of hydrologic conditions. These are both used to characterize the time series and as inputs to clustering algorithms.

The water levels show the sequence of unsaturated, saturated and flooded conditions at different heights and depths. This is especially important for inferring redox conditions for GHG models. The first differences of the time series are the one-day lag changes in the water level in the pani-pipes. They show the speed at which water level changes. This is related to the “flashiness” of the hydrology.

The 17 descriptors are meant to represent and summarize the hydrologic behaviour of field water level to be easily associated with management and models outcomes, for example when estimating GHG emissions using process-based models. The descriptors are of three general types, according to the aspect of the hydrology that they

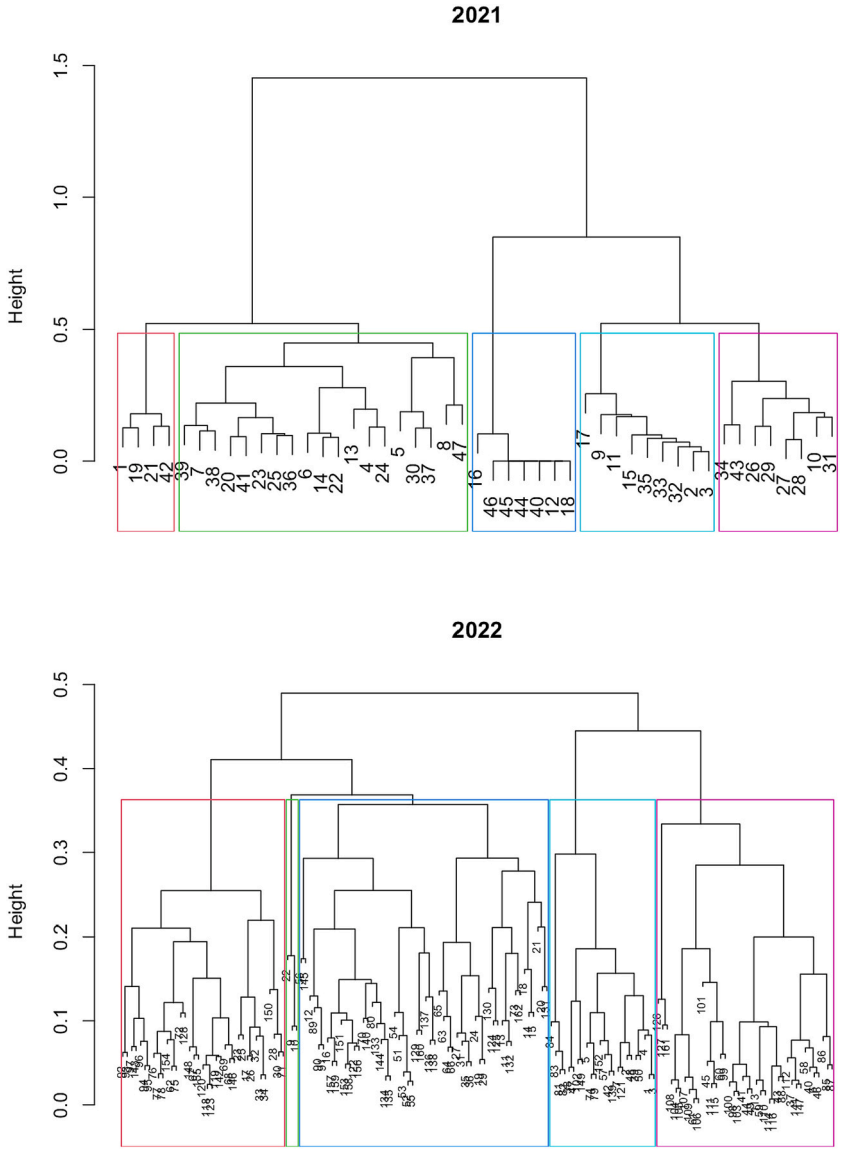


Fig. 8. Hierarchical clustering of descriptors, 2021 (top), 2022 (bottom).

characterize.

1. Flood periods: (1) percent of days flooded, i.e., water is above the soil surface (days.flood.pct, %); (2) percent of days of soil flooded at depths of 0–5 cm (flood.0.5.cm, %); (3) at depths of 5 to 10 cm (flood.5.10.cm, %); (4) at deeper than 10 cm (flood.10.cm, %); (5) Percent of days of flood events with duration less than two days (flood.events.one.day, %); (6) same, within a week (flood.events.one.week, %); (7) same, within one month (flood.events.one.month, %); (8) same, more than one month (flood.events.more.month, %).
2. Drainage periods (water level ≤ 0): (9) percent days of with the water level between −5–0 cm (flood.5.0.cm, %); (10) same, −10–-5 cm (flood.10.5.cm, %); (11) same, deeper than −10 cm (flood.10.minus.cm, %); (12) average duration of soil drainage at −5–0 cm (avg.duration.dry.5.0, days); (13) same, −10–-5 cm (avg.duration.dry.10.5, days); (14) same, deeper than −10 cm (avg.duration.dry.10.minus, days).
3. Transition between potentially saturated/unsaturated conditions: (15) number of flood events (flood.events, count); (16) number of days when water level drops from 5 cm to −5 cm (days.redox, count); (17) number of days when the water level drops more than 15 cm (days.quick.drain, count). Although the pipes only show the water level, and the saturation status and redox potential is not an instantateous response to water level, this approximation is the closes we can come with our information to estimating saturated/unsaturated conditions.

Several of these descriptors have special significance for GHG modelling and land management. For example, descriptor (16) represents the transition from potentially reducing to potentially oxidizing conditions near the soil surface, while descriptor (17) represents quick drainage. However, all are useful for clustering into groups from which representative sites can be selected for GHG models.

Tables 2 (2021) and 3 (2022) show the summary statistics for the descriptors.

4.2. *Clustering*

Both the time series (concatenated first differences and absolute water levels) and their descriptors were used to cluster the pani-pipes into functional groups. Before attempting any clustering, we checked the clustering tendency with the Hopkins statistic calculated with the

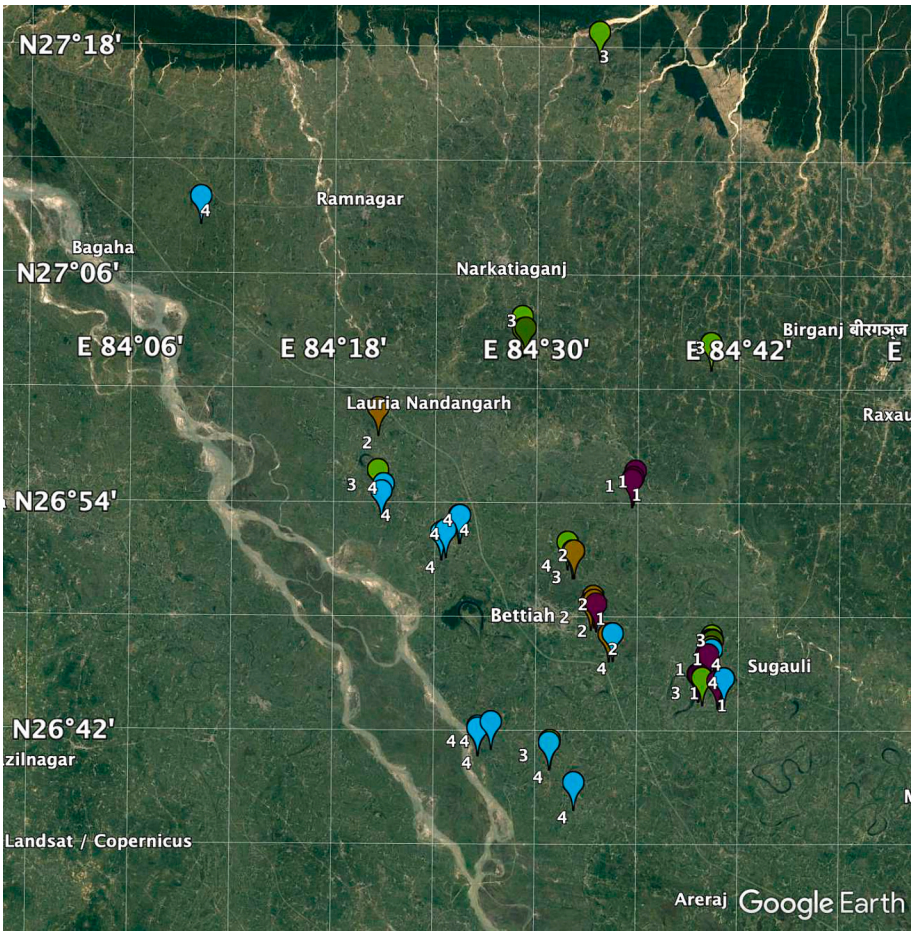


Fig. 9. Location of pani-pipes labelled by cluster number (k-means, time series), 2021. Representative locations shown with yellow push-pin.

hopkins R package, and only proceeded if this indicated that clustering is sensible. We clustered using two approaches. The first was k-means clustering, implemented by comparing the kmeans function of the stats package with the default Hartigan-Wong algorithm and 24 random starting positions. An optimum number of clusters was determined by comparing the Beale, silhouette, elbow and gap statistic indices computed by the NbClust function of the NbClust R package (Charrad et al., 2014). The relative importance of the descriptors for k-means clustering was evaluated with the FeatureImpCluster R package, which implements permutation feature importance as described by Molnar (2022).

The second approach was hierarchical clustering implemented by the hclust function of the stats R package to produce a dendrogram, with the "Ward.D2" linkage method, which produces compact, spherical clusters. For this, we used the number of clusters determined by comparing the indices for k-means clusters. The dendrograms reveal the distance between clusters, but also show potential clusters at coarser and finer levels of detail.

Cluster assignments were compared with cross-classification tables, and the agreement between methods or years was quantified as the average of the weighted (by total in a cluster) matches to the dominant cluster of the other classification (Eq. 1):

$$\frac{1}{2}\left\{\sum_i \left(max x_{ij} / x_{i+}\right) * \left(x_{i+} / x_{++}\right) + \left(\sum_j max y_{ji} \Big/ y_{j+}\right) * \left(y_{j+} \Big/ y_{++}\right)\right\} \quad (1)$$

where x_{ij}, y_{ji} is an individual matrix entry in row i and column j of the matrix, x_{i+}, y_{j+} are row and column sums, respectively, and $x_{++} = y_{++}$ is the total number of observations. This clustering quality index is 1.0 for perfect agreement (whether the cluster numbers are the same), and from $\approx 0.35 \pm 0.05$ for a random assignment to clusters, indicating no correspondence.

Representative pani-pipes were identified for each cluster, for both methods and both data sources, by finding the closest observation, in Euclidean feature space, to the k-means cluster centroids.

4.3. Comparison with landscape descriptors and soil properties

The clusters were used as one margins (row or column) of cross-classification tables, with the other margin being landscape descriptors (farmer's perception of landscape position) or the rice management system (DSR and TPR). This was quantified by the clustering quality index developed for comparing clusters (§4.2). The cluster assignments were also used as factors in one-way analysis of variance (ANOVA) and means separation of soil properties. This reveals to what degree soil properties are related to the empirically-derived clusters based on hydrologic behaviour.

A more general question is which factors (e.g., soil, groundwater resources, terrain elevation, irrigation management) can best explain the differences in hydrologic behaviour between clusters.

For this, random forest models were built to predict the empirically-derived cluster number from these factors. Management choices (only for 2022) included as predictors in the models were planting method (direct-seeded or transplanted), number of irrigations, timing of irrigations (early vegetative, mid vegetative, flowering and grain filling), reason for irrigating (visible crop stress, soil cracking, disappearance of flood water, crop growth stage), depth of tube well (shallow or deep), pump type and power. Note that these factors are not necessarily direct

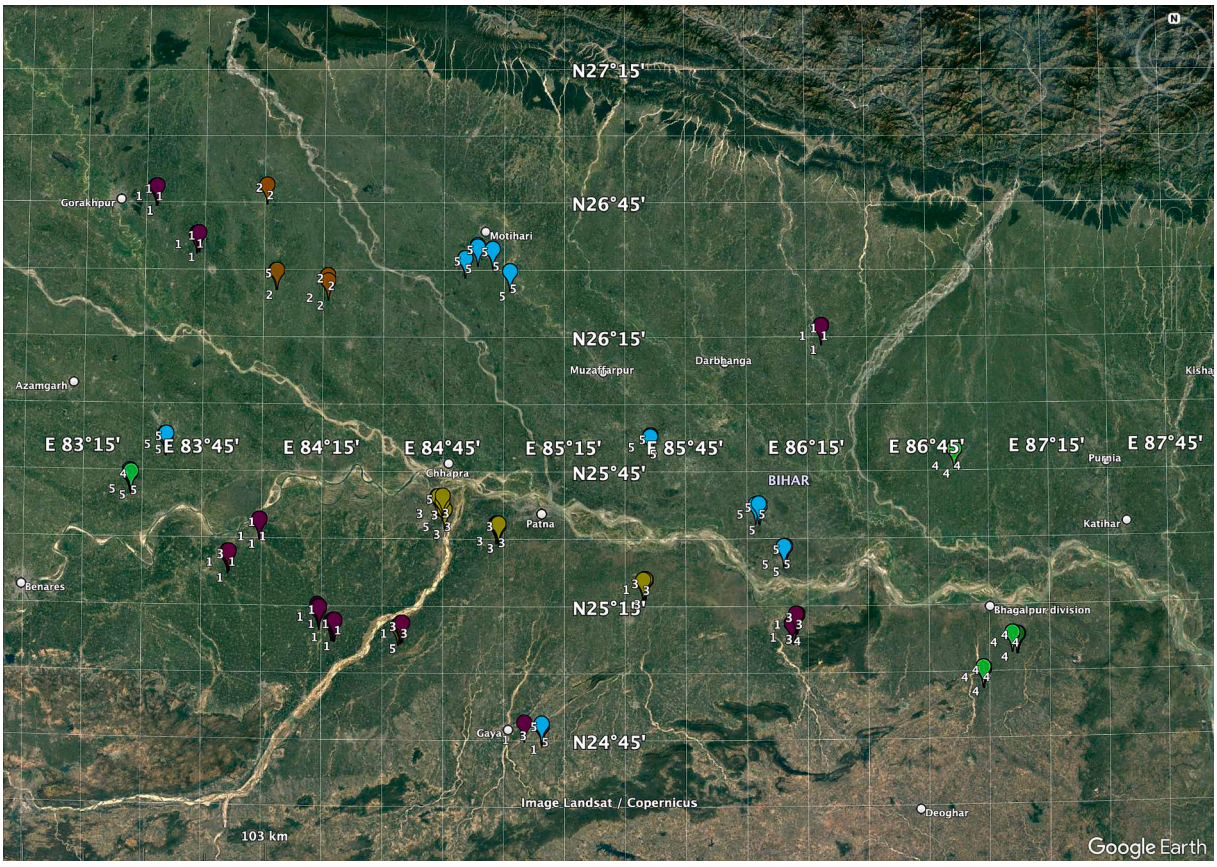


Fig. 10. Location of pani-pipes labelled by cluster number (k-means, time series), 2022. Representative locations shown with yellow push-pin.

causes of a site's being in a given cluster – the random forest model is predictive, not explanatory. But their importance in the fitted model may be useful for interpretation. Other predictors were pre- and post-monsoon water levels and their difference (seasonal change) in the nearest groundwater well, terrain elevation, and annual rainfall at the nearest rain gauge. Certainly, other factors could be important predictors, notably subsoil properties that affect water movement. Unfortunately these were not included in the experimental design.

Random forest classification models were built with the ranger R package (Wright and Ziegler, 2017) and optimized by recursive feature elimination with the caret package (Kuhn, 2008). Random forest probability models were also built, to evaluate the uncertainty of class assignment. The sets of class probabilities for each observation were quantified by the Confusion Index and Shannon entropy. The Confusion Index (Burrough et al., 1997) (Eq. 2) is defined as:

$$\mathrm{CI} = 1 - \left\{ \mu_{\max} - \mu_{(max-1)} \right\} \tag{2}$$

where $\mu_{\max}$ is the probability of the most probable class, and μ_{max-1} is the probability of the second most probable class. It shows how well the classification is separated from its alternative. Shannon entropy (Eq. 3) shows the information uncertainty. For a variable z with n classes, each of which has estimated proportion $\widehat{\pi}(z_i)$:

$$H_z = -\sum_{i=1}^{n} \widehat{\pi}(z_i) \cdot log_n \widehat{\pi}(z_i) \tag{3}$$

The reason to use base-n logarithms is so that 0 represents no uncertainty, and 1 maximum (Kempen et al., 2009).

Paired (within 200 m) DSR and TPR plots were identified and their time-series of water levels compared visually.

5. Results

5.1. Clusters and their hydrologic regimes

5.1.1. Water levels

Using the water levels and their first differences, in both 2021 and 2022 the optimum number of clusters using k-means was two; however, it is clear that there are more response units on the agricultural landscape. From the plots of number of clusters vs. indices we found that four (2021) and five (2022) clusters were quite close to the optimum, and so continued the analysis with these groupings. The various optimization criteria did not give consistent recommendations for optimization, showing that clustering was not clear-cut. The model efficiency coefficient (MEC) of k-means clustering on time-series were 0.72 (2021, four clusters) and 0.41 (2022, five clusters). This shows that 2022 (dry conditions) was a less consistent year with more poorly-defined clusters, and that in 2021 (wet conditions) fewer clusters were indicated.

Figs. 1 and 2 show the first differences of the time-series of water levels for 2021 and 2022 at the cluster centroids, followed by the absolute levels. There are clear differences between the cluster centroids in both years.

In 2021 (wet year) Cluster 2 had standing water at least 15 cm deep for the entire monitoring period. Cluster 1 was completely flooded, or nearly so, through the early part of the growing season, then had a 20-day dry-down phase, followed by one more flooding and final dry-down towards harvest. Cluster 3 was similar to Cluster 1 but with shallower flooding and earlier first dry-down. The water levels in Cluster 4 had many rapid changes from a few cm above to a few cm below the soil surface in the early part of the growing season, and then were quite dry except for a single event matching the late-season spikes of Clusters 1 and 3.

In 2022 (dry year) the most interesting cluster is 2, with rapid

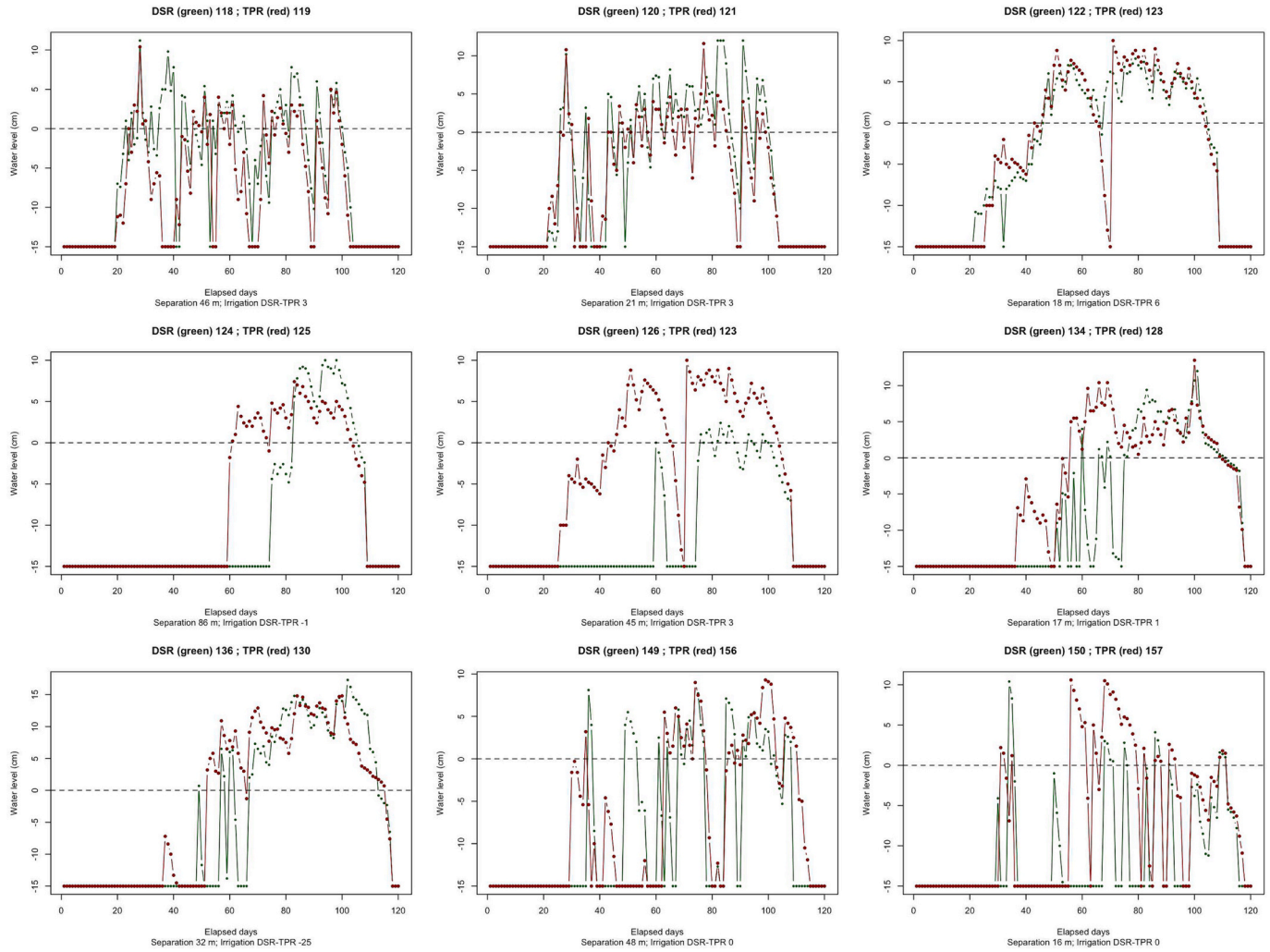


Fig. 11. Paired DSR (green) / TPR (red) time-series of water levels at pani-pipes.

transitions between quite wet and quite dry conditions. This cluster is quite "flashy", with sharp short-term changes in water levels. Cluster 1 was mostly dry, with only small fluctuations between −15 and − 10 cm. Cluster 3, 4 and 5 had similar temporal patterns: dry early, then water reaching or somewhat exceeding the soil surface, followed by a late-season dry-down. These three differ in the timing and magnitude of the wetter periods. All clusters had an initial dry period.

Another visualization is given by the hierarchical clustering of the time-series shown in Figs. 3 (2021) and 4 (2022). This shows the hydrological response of all of the sites and the variability within each cluster. Reading the vertical axis (all sites) at any given day on the horizontal axis reveals events that affect different numbers of sites. Examining the lengths of dendrogram stems shows how different are the clusters, and the members within each cluster. For example, in 2021 Cluster 2 (as identified by k-means) is uniform: always completely flooded. However, Cluster 1 has some variability in the timing and magnitude of the dry-down, although the overall pattern is consistent. There is not much difference between members, as evidenced by the short stems. In 2022 the differentiation between the clusters is not so apparent, and there is more fine-scale variability within them. However, Cluster 2 is clearly differentiated by its short-term fluctuations ("flashy" behaviour).

5.1.2. Descriptors

Many of the 17 hydrologic descriptors were correlated (Fig. 5), but the pattern of correlation differed among years. While descriptors of flood duration were positively correlated with the total days flooded in 2021, this correlation was only observed for flood duration of one day in 2022. The lack of consistent correlation patterns between descriptors is associated with the occurrence of a given descriptor. For instance, flood durations of one month or longer were less frequent in 2022 than floods lasting between one day and one week. Hence, the positive correlation of floods is linked to the most common flood duration. The direction and strength of correlations are both associated with how well the descriptor can describe the hydrologic behaviour characterizing the season.

For the k-means clustering using hydrologic descriptors as input, the several optimization measures were not consistent. A reasonable compromise was five clusters for 2021 (as opposed to four from the clustering based on time-series), as well as 2022 (in agreement with the time-series). The MEC for these was 0.82 (2021) and 0.67 (2022), again showing that clusters were better-defined in the wet year, 2021.

Fig. 6 shows the positions of each observation in the space spanned by the first two standardized principal components (PC) of the descriptors, along with the convex hulls of the cluster groups. In 2021 the first two PCs account for 60.5% of the total information in the descriptors and the clusters are mostly well-separated in this space; in 2022 this is only 52.3%. and there is more overlap of the convex hulls. This again shows that the cluster structure is stronger in the wetter year.

The relative feature importance for k-means clustering is shown in Fig. 7. For both years, the number of flood events of more than one month and week were the most important descriptors, but in reverse order: the longer period for 2021 (wet year) and the shorter for 2020

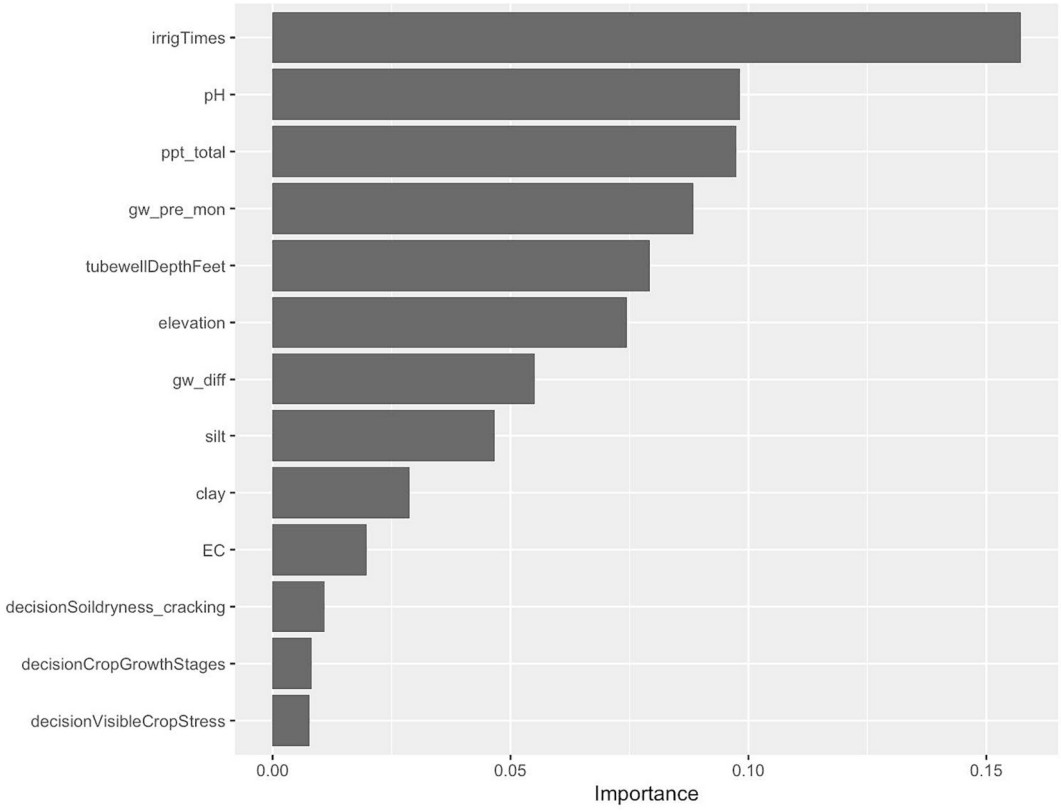


Fig. 12. Variable importance of 2022 random forest model to predict cluster membership, hierarchical clustering on time-series.

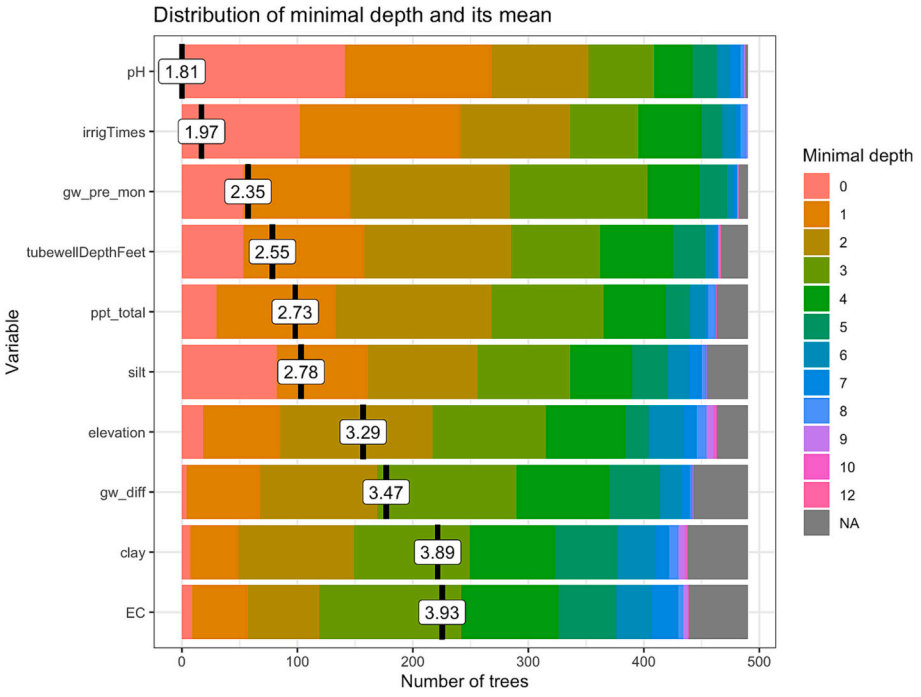


Fig. 13. Levels in 2022 random forest trees at which predictors are used, hierarchical clustering on time-series.

(dry year). After these two there is a clear difference between the years: in the wet year the number of days flooded to at least 10 cm and the proportion of flooded days but in the dry year one-day flood events.

Descriptors that summarize flooding depth and timing proved to be the most relevant for k-means clustering. The key change is that non-averaged descriptors not only exhibit higher representation during the season, but are also more sensitive across clusters. On the contrary, average duration of drained periods and transitions between saturated and unsaturated conditions are mostly irrelevant for clustering in both years. In both years many descriptors are irrelevant, and several are

Table 4
K-means (rows) vs. hierachical (columns) clustering on time series, 2021 (left), 2022 (right). Quality of match 0.936 (2021), 0.888 (2022).

	1	2	3	4	5	1	2	3	4	5
1	0	12	0	0		10	0	36	0	0
2	0	0	0	6		0	0	0	0	7
3	1	1	11	0		32	1	0	0	0
4	15	0	1	0		0	1	0	28	0
5						2	39	0	3	1

Table 5
K-means (rows) vs. hierachical (columns) clustering on descriptors, 2021 (left), 2022 (right). Quality of match 0.745 (2021), 0.631 (2022).

	1	2	3	4	5	1	2	3	4	5
1	4	0	6	0	0	8	3	0	26	37
2	0	0	0	0	7	0	7	0	0	0
3	0	0	7	0	0	0	3	0	19	0
4	0	9	0	3	0	0	0	3	11	0
5	0	0	6	5	0	16	27	0	0	0

Table 6
K-means clustering on time series (rows) vs. descriptors (columns), 2021 (left), 2022 (right). Quality of match 0.809 (2021), 0.644 (2022).

	1	2	3	4	5	1	2	3	4	5
1	0	1	0	11	0	28	0	9	9	0
2	0	6	0	0	0	0	0	0	0	7
3	0	0	1	1	11	25	0	1	1	6
4	10	0	6	0	0	13	0	12	4	0
5						8	7	0	0	30

Table 7
Hierarchical clustering on time series (rows) vs. descriptors (columns), 2021 (left), 2022 (right). Quality of match 0.702 (2021), 0.613 (2022).

	1	2	3	4	5	1	2	3	4	5
1	4	0	12	0	0	11	0	0	14	19
2	0	8	0	4	1	8	33	0	0	0
3	0	1	7	4	0	0	0	3	18	15
4	0	0	0	0	6	1	3	0	24	3
5						4	4	0	0	0

Table 8
Farmer perceptions vs. k-means clusters from descriptors, 2021 (left), 2022 (right). Quality of match 0.468 (2021), 0.621 (2022).

	1	2	3	4	5	1	2	3	4	5
Lowland	0	7	1	11	1	11	0	4	1	1
Medium	9	0	6	1	10	16	0	4	2	3
Upland	1	0	0	0	0	9	0	0	3	4

Table 9
2022: Rice planting method vs. k-means clusters from descriptors, quality of match 0.462.

	1	2	3	4	5
DSR	31	3	7	3	28
TPR	43	4	15	11	15

irrelevant in both years: days of transition to reducing conditions, days of quick drainage, and average duration of dry periods. Some of this is due to the strong correlation between descriptors, so if one is permuted, another one can replace it in the clustering. In addition, some descriptions are indeed well-correlated with water dynamics but do not vary across clusters in a given year, so are not important in clustering. Examples are descriptors of long-term flood duration in sites with overall dry conditions, or the number of drainage days at deeper soil layers in predominantly saturated conditions.

Another visualization is given by the of hierarchical clustering of the descriptors are shown in Fig. 8. As with the hierarchical clustering from time-series, 2021 shows a much stronger differentiation between clusters than 2022, and much more consistency within clusters, as shown by the shorter stems.

5.1.3. *Spatial distribution of clusters*

Figs. 9 (2021) and 10 (2022) show the locations over the landscape of each observation well and its cluster group from k-means clustering on time series. There is some spatial differentiation between clusters, especially for 2022, but also nearby locations in different clusters.

5.1.4. *Matching field-based clusters*

A key question is how well do the clusters in each of the two years match (1) between methods with the same source information (time-series vs. descriptors), (2) between different source information, for the two methods. This reveals whether there are consistent HRT between and within years.

Table 4 shows the cross-classification between clustering on time-series by k-means and hierarchical clustering for both years. The agreement between methods in 2021 (wet year) is very good, in 2022 somewhat less so but still good.

Table 5 shows the cross-classification between clustering on descriptors by k-means and hierarchical clustering for both years. The agreement between methods is again better in 2021 than 2022, but the agreements are less than when clustering by time-series.

Table 6 (2021) shows the cross-classification of the k-means clustering using time-series and descriptors, and Table 7 shows the same for hierarchical clustering. The k-means clusters based on the two information sources is again better in 2021 than 2022, but in neither case is the match particularly good. A similar result is found for hierarchical clustering. This shows that the information contained in the two sources are different, and the choice between them depends on how the analyst conceptualizes the soil hydrology: directly from a time-series or as descriptors that can be interpreted in terms of hydrologic behaviour related to a specific land management objective.

5.1.5. *Comparison with soil properties*

For both clustering methods (k-means and hierarchical clustering) and both data sources (time-series and descriptors), there was quite a poor separation of soil properties by cluster, as measured by the adjusted R^2 of a one-way ANOVA, despite the fairly wide range of the soil properties at the observation sites. For 2021, only soil organic matter concentration (SOC, R^2 0.13–0.16), electrical conductivity (EC, R^2 0.10–0.18) and pH in water (pH, R^2 0.04–0.11) were even weakly separated by cluster; for 2022 again pH (R^2 0.24–0.35) and EC (R^2 0.11–0.14) but not SOC, as well as silt concentration (R^2 0.11–0.29), sand (R^2 0.05–0.09), and clay (R^2 0.05–0.15). Differences in soil texture were not associated with hydrologic behaviour in 2021 (the wet year) but were in 2022 (the dry year), showing that soil texture does have some influence on soil hydrology under drier conditions, as the finer textures are associated with greater water-holding capacity, but this is not important when water is abundant. This shows that the hydrologic behaviour in this landscape is only weakly related to soil properties.

5.1.6. *Comparison with landscape descriptors*

For both clustering methods (k-means and hierarchical clustering) and both data sources (time-series and descriptors), there was quite a poor relation with farmer perceptions of landscape position. This was assessed by the cross-classification clustering quality index (§4.2). For

2021, the index ranged from 0.49 to 0.60 (based on all 43 observations), for 2022 from 0.46 to 0.62 (based on 58 of the 160 observations). The closest matches for the two years are shown in Table 8. No clear relation can be seen either within a cluster or a farmer-perceived class, except in the wet year (2021) when one cluster represented the permanently-flooded sites, is all farmer-perceived "lowland"; however this includes only 11 of the 20 perceived "lowlands".

In 2022 the rice planting method was recorded for each site. There was almost a random relation between the clusters and the planting method; the cross-classification quality index ranged from 0.34 to 0.46 for the different clustering methods. Table 9 shows the closest match, for k-means clustering from descriptors. There is some differentiation: clusters 1, 3 and 4 have more TPR, whereas cluster 5 has more DSR.

5.1.7. Paired DSR/TPR plots

Forty DSR plots had a TPR plot within 200 m. These were considered as paired plots. The number of irrigation events differed somewhat among pairs, from −6 to +6 (with one outlier due to river irrigation) but with 70% of the pairs within 2 irrigations. Despite the close spacing, substantial differences in soil properties were found between the pairs: the interquartile ranges were − 7.5–10% (sand), −10–5% (silt), −2.5–3.1% (clay), −0.4–1.3 pH, 0.04–0.49% (OC). Fig. 11 shows the paired time-series of water level for six of the pairs. In several of these (e.g., locations 124/125) management (perhaps soil puddling in TPR) control much of the hydrology, whereas in others (e.g., locations 150/157) the two series are similar, implying that hydrology is controlled by rainfall and/or irrigation.

5.2. Predicting soil hydrology with soil, agronomic management, and landscape characteristics

The optimized random forest models for 2021 included only two predictors: precipitation and pre-monsoon groundwater level. No soil properties were selected. This is likely because the sites were flooded for most of the period and so the water supply (precipation and groundwater) was sufficient to explain the response. For clusters based on descriptors, both k-means and hierarchical clustering gave OOB classification errors from 60% (k-means) to 68% (hierarchical), i.e., almost no skill. The models for the clusters based on the time-series of water levels and their first differences were better, 36% (hierarchical) to 40% (k-means) OOB error, still very poor predictive power. Total precipitation was the most important predictor, closely followed by pre-monsoon groundwater level.

The optimized random forest models for 2022 included (1) landscape features: total rainfall, pre-monsoon groundwater level, within-season difference in groundwater level, tubewell depth, and terrain elevation; (2) soil properties: pH, EC, clay and silt concentration; (3) management: number of irrigations and whether irrigation was (partly) decided based on crop stress, visible soil dryness, or crop growth stage. These models had quite uneven success, depending on the clustering method used to establish the clusters. For clusters based on descriptors, both k-means and hierarchical clustering gave OOB classification errors in around 36% of cases. The models for the clusters based on the time-series of water levels and their first differences were much better, around 23% OOB classification error for k-means but only 15% for hierarchical clustering. Variable importance for the latter model is shown in Fig. 12 and the level at which predictors are used in the various trees of the random forest in Fig. 13 The number of irrigations and the total rainfall have a large influence in this very dry year. Groundwater level and the seasonal difference are also important. The high importance of pH is likely because it can substitute for silt concentration.

The probability random classification forests for 2021 and 2022 revealed quite high uncertainty in assignment to hierarchical clusters based on time series. For 2021, maximum probabilities ranged from 0.35 to 0.85, median 0.56. The Confusion Index ranged from 0.21 (moderate difference in probability between the two most probable classes) to 0.95 (the highest-probability class almost certain), median 0.73. Normalized Shannon entropy ranged from 0.36 (moderate uncertainty) to 0.82(high uncertainty), median 0.62. For 2022, maximum probabilities ranged from 0.39 to 0.98, median 0.86. The Confusion Index ranged from 0.03 (almost no difference in probability between the two most probable classes) to 0.96 (the highest-probability class almost certain), median 0.23 Normalized Shannon entropy ranged from 0.08 (almost no uncertainty) to 0.85 (high uncertainty), median 0.35.

Thus, for both years there was a wide range in classification uncertainties: some observations were well-predicted, others quite poorly. This shows that clusters do not cleanly divide the observations into HRT, consistent with other evidence to this effect.

6. Conclusion

Hydrologic response at the pani-pipes could be fairly successfully clustered, both by examining the time-series (water levels) and their first differences (changes in water levels) and a set of descriptors of hydrologic response. K-means clustering was able to establish sets of similar behaviour. Hierarchical clustering also did this, but in addition showed the distance in feature space between and within clusters. Clustering was more successful in a wet year (2021) than in a dry year (2022). However, the two years could not be directly compared, because the observation sites were not the same.

The spatial-temporal distribution of water levels in the field responds to various drivers. While the distribution of flooding in 2021 can be attributed to natural factors like rainfall and horizontal flows, in 2022 the observed water levels resulted from a combination of both natural and anthropogenic drivers. During 2022 the frequency of floods might correspond to several irrigation criteria, which may not necessarily align with large-scale patterns. This suggests that defining HRT must be per-year. However, the inconsistent results with the various clustering methods, even within a year, suggests that there may be no consistent way to define HRT in this landscape.

We had hoped to establish a grouping that could be extrapolated across years and across the eastern IGP. The inconsistency between the years, and the disappointing results of the attempt to identify predictive landscape factors, show that this is, so far, not feasible. It may be that subsoil properties, not accounted for in the predictive factors set, would result in a more successful model.

Clustering was only moderately consistent among methods and information sources. Part of the problem is when the water level is below ground, but we have no estimate of the water content in the "dry" section above the water table. Depending on the porosity and structure, and the time since drainage, the water content can be anywhere from almost saturated to the permanent wilting point. Thus, the indicators for the drained conditions at the various levels in the profile only roughly track the measured water levels.

Clustering showed very weak relations with farmer perceptions of landscape position and a weak relation with rice planting method.

Clusters for 2022 based on time-series could be fairly well predicted by optimized random forest models. The most important predictors were the number of irrigations, seasonal precipitation, pre-monsoon groundwater levels, seasonal groundwater level change, and pH, this latter as a surrogate for landscape position and other soil properties. This shows the complex relation of soil hydrology with landscape position and land management.

The use of a set of descriptors to summarize the time series was here specific to presumed relation with GHG emissions. Another set of descriptors could be developed to reflect other land use issues affected by soil hydrology, e.g., wet-up at the beginning of the *kharif* season to allow rice planting, or water levels related to weed management.

This study was opportunistic, using sets of field observations that had not been designed to characterize hydrology. Sites were selected for agronomic and landscape ecology interest, not as transects to characterize hydrology. Further, only two years were characterized and the

observation areas differed among years, so that the difference between years may be partially confounded with the difference in location. Still, these preliminary results suggest that a study fit for purpose would be interesting. A proper characterization of soil hydrology to land management in the IGP is a worthy challenge, to which this study is a first step.

Funding sources

This work was supported, in whole or in part, by the Bill & Melinda Gates Foundation [INV-029117]. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic Licence has already been assigned to the Author Accepted Manuscript version that might arise from this submission.

CRediT authorship contribution statement

D.G. Rossiter: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Laura Arenas-Calle:** Writing – review & editing, Formal analysis, Conceptualization. **Anton Urfels:** Writing – review & editing, Conceptualization. **Harishankar Nayak:** Writing – review & editing, Methodology, Conceptualization. **Sonam Sherpa:** Writing – review & editing, Supervision, Project administration, Methodology. **Andrew McDonald:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

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

Andrew McDonald reports financial support was provided by Bill & Melinda Gates Foundation. 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.

Data availability

Data will be made available on request.

References

Allaire, J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., Iannone, R., 2023. rmarkdown: Dynamic Documents for R. URL. https://github.com/rstudio/rmarkdown.r package version 2.21.

Amazon Web Services, 2023. Terrain Tiles - Registry of Open Data on AWS. https://registry.opendata.aws/terrain-tiles/.

Balwinder-Singh, McDonald, A.J., Kumar, V., Poonia, S.P., Srivastava, A.K., Malik, R.K., 2019. Taking the climate risk out of transplanted and direct seeded rice: insights from dynamic simulation in Eastern India. Field Crop Res. 239, 92–103. https://doi.org/10.1016/j.fcr.2019.05.014.

Bo, Y., Jaegermeyr, J., Yin, Z., Jiang, Y., Xu, J., Liang, H., Zhou, F., 2022. Global benefits of non-continuous flooding to reduce greenhouse gases and irrigation water use without rice yield penalty. Glob. Chang. Biol. 28, 3636–3650. https://doi.org/10.1111/gcb.16132.

Bonsor, H.C., MacDonald, A.M., Ahmed, K.M., Burgess, W.G., Basharat, M., Calow, R.C., Dixit, A., Foster, S.S.D., Gopal, K., Lapworth, D.J., Moench, M., Mukherjee, A., Rao, M.S., Shamsudduha, M., Smith, L., Taylor, R.G., Tucker, J., van Steenbergen, F., Yadav, S.K., Zahid, A., 2017. Hydrogeological typologies of the Indo-Gangetic basin alluvial aquifer, South Asia. Hydrogeol. J. 25, 1377–1406. https://doi.org/10.1007/s10040-017-1550-z.

Boorman, D.B., Hollis, J.M., Lilly, A., 1995. Hydrology of Soil Types: A Hydrologically-Based Classification of the Soils of the United Kingdom. Technical Report. UK Institute of Hydrology.

Burrough, P., van Gaans, P., Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77, 115–135. https://doi.org/10.1016/S0016-7061(97)00018-9.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., 2014. Nbclust: an R package for determining the relevant number of clusters in a data set. J. Stat. Softw. 61, 1–36. https://doi.org/10.18637/jss.v061.i06.pdf.

CIMMYT, 2017. The Cereal Systems Initiative for South Asia (CSISA). https://csisa.org.

Deng, J., Zhou, Z., Zhu, B., Zheng, X., Li, C., Wang, X., Jian, Z., 2011. Modeling nitrogen loading in a small watershed in Southwest China using a DNDC model with hydrological enhancements. Biogeosciences 8, 2999–3009. https://doi.org/10.5194/bg-8-2999-2011.

Faouzi, E., Arioua, A., Namous, M., Barakat, A., Mosaid, H., Ismaili, M., Eloudi, H., Hanadé Houmma, I., 2023. Spatial mapping of hydrologic soil groups using machine learning in the Mediterranean region. CATENA 232, 107364. https://doi.org/10.1016/j.catena.2023.107364.

Garen, D.C., Moore, D.S., 2005. Curve number hydrology in water quality modeling: uses, abuses and future directions. J. Am. Water Resour. Assoc. 41, 377–388. https://doi.org/10.1111/j.1752-1688.2005.tb03742.x.

Government of India, 2022. Open Government Data (OGD) Platform India. https://data.gov.in/catalog/rainfall-india.

Jain, V., Sinha, R., 2003. River systems in the Gangetic plains and their comparison with the Siwaliks: a review. Curr. Sci. 84, 1025–1033 (arXiv:24107664).

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. Geoderma 151, 311–326. https://doi.org/10.1016/j.geoderma.2009.04.023.

Kraus, D., Weller, S., Klatt, S., Haas, E., Wassmann, R., Kiese, R., Butterbach-Bahl, K., 2015. A new LandscapeDNDC biogeochemical module to predict CH4 and N2O emissions from lowland rice and upland cropping systems. Plant Soil 386, 125–149. https://doi.org/10.1007/s11104-014-2255-x.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26. https://doi.org/10.18637/jss.v028.i05.

Lilly, A., Boorman, D.B., Hollis, J.M., 1998. The development of a hydrological classification of UK soils and the inherent scale changes. Nutr. Cycl. Agroecosyst. 50, 299–302. https://doi.org/10.1023/A:1009765000837.

McDonald, A.J., Riha, S.J., Duxbury, J.M., Steenhuis, T.S., Lauren, J.G., 2006. Water balance and rice growth responses to direct seeding, deep tillage, and landscape placement: findings from a valley terrace in Nepal. Field Crop Res. 95, 367–382. https://doi.org/10.1016/j.fcr.2005.04.006.

McDonald, A.J., Balwinder-Singh, Keil, A., Srivastava, A., Craufurd, P., Kishore, A., Kumar, V., Paudel, G., Singh, S., Singh, A.K., Sohane, R.K., Malik, R.K., 2022. Time management governs climate resilience and productivity in the coupled rice–wheat cropping systems of eastern India. Nat. Food 3, 542–551. https://doi.org/10.1038/s43016-022-00549-0.

McDonald, A.J., Malik, R.K., Ajay, A., Craufurd, P., Dubey, S., Gautam, U., Karki, S., Kishore, A., Krupnik, T.J., Kumar, V., Mkondiwa, M., Nayak, H.S., Parihar, C.M., Paudel, G., Peramaiyan, P., Pundir, A., Poonia, S., Samaddar, A., Sherpa, S., Singh, B., Singh, S., Urfels, A., Chellattan Veettil, P., Pathak, H., Singh, A.K., 2023. Bigger Data from Landscape-Scale Crop Assessment Surveys Empowers Sustainability Transitions. https://doi.org/10.2139/ssrn.4511866.

Miro, B., Ismail, A., 2013. Tolerance of anaerobic conditions caused by flooding during germination and early growth in rice (Oryza sativa L.). Front. Plant Sci. 4 https://doi.org/10.3389/fpls.2013.00269.

Molnar, C., 2022. Interpretable Machine Learning, Second edition. Leanpub.

Park, S.J., van de Giesen, N., 2004. Soil-landscape delineation to define spatial sampling domains for hillslope hydrology. J. Hydrol. 295, 28–46. https://doi.org/10.1016/j.jhydrol.2004.02.022.

Quisenberry, V.L., Smith, B.R., Phillips, R.E., Scott, H.D., Nortcliff, S., 1993. A soil classification system for describing water and chemical transport. Soil Sci. 156, 306. https://doi.org/10.1097/00010694-199311000-00003.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL. https://www.R-project.org.

Ross, C.W., Prihodko, L., Anchang, J.Y., Kumar, S.S., Ji, W., Hanan, N.P., 2018. Global hydrologic soil groups (HYSOGs250m) for curve number-based runoff modeling. ORNL DAAC. https://doi.org/10.3334/ORNLDAAC/1566.

Semeniuk, C.A., Semeniuk, V., 2018. Wetland classification: Hydrogeomorphic system. In: Finlayson, C.M., Everard, M., Irvine, K., McInnes, R.J., Middleton, B.A., van Dam, A.A., Davidson, N.C. (Eds.), The Wetland Book: I: Structure and Function, Management, and Methods. Springer Netherlands, Dordrecht, pp. 1483–1489. https://doi.org/10.1007/978-90-481-9659-3_331.

Sinha, R., Jain, V., Babu, G.P., Ghosh, S., 2005. Geomorphic characterization and diversity of the fluvial systems of the Gangetic Plains. Geomorphology 70, 207–225. https://doi.org/10.1016/j.geomorph.2005.02.006.

Timsina, J., Connor, D.J., 2001. Productivity and management of rice–wheat cropping systems: issues and challenges. Field Crop Res. 69, 93–132. https://doi.org/10.1016/S0378-4290(00)00143-X.

Urfels, A., McDonald, A.J., van Halsema, G., Struik, P.C., Kumar, P., Malik, R.K., Poonia, S.P., Balwinder-Singh, Singh, D.K., Singh, M., Krupnik, T.J., 2021. Social-ecological analysis of timely rice planting in Eastern India. Agron. Sustain. Dev. 41, 14. https://doi.org/10.1007/s13593-021-00668-1.

USDA Natural Resources Conservation Service, 2020. National Engineering Handbook: Part 630 - Hydrology; Chapter 7 Hydrologic Soil Groups. H_210_NEH_630.

Wright, M.N., Ziegler, A., 2017. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77, 1–17. https://doi.org/10.18637/jss.v077.i01.

Yin, S., Zhang, X., Lyu, J., Zhi, Y., Chen, F., Wang, L., Liu, C., Zhou, S., 2020. Carbon sequestration and emissions mitigation in paddy fields based on the DNDC model: a review. Artif. Intell. Agric. 4, 140–149. https://doi.org/10.1016/j.aiia.2020.07.002.