University of Reading

Using interpretable gradient-boosted decision-tree ensembles to uncover novel dynamical relationships governing monsoon low-pressure systems

Article

Published Version

Creative Commons: Attribution 4.0 (CC-BY)

Open Access

Hunt, K. M. R. ORCID: https://orcid.org/0000-0003-1480-3755 and Turner, A. G. ORCID: https://orcid.org/0000-0002-0642-6876 (2024) Using interpretable gradient-boosted decision-tree ensembles to uncover novel dynamical relationships governing monsoon low-pressure systems. Quarterly Journal of the Royal Meteorological Society, 150 (758). pp. 1-24. ISSN 1477-870X doi: 10.1002/qj.4582 Available at https://centaur.reading.ac.uk/113085/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1002/qj.4582

Publisher: Royal Meteorological Society

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in

the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Check for updates

Received: 6 February 2023 | Revised: 4 July 2023 | Accepted: 28 August 2023

DOI: 10.1002/qj.4582

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Using interpretable gradient-boosted decision-tree ensembles to uncover novel dynamical relationships governing monsoon low-pressure systems

Kieran M. R. Hunt[1,2] | **Andrew G. Turner**[1,2]

[1]Department of Meteorology, University of Reading, Reading, United Kingdom

[2]National Centre for Atmospheric Science, University of Reading, Reading, United Kingdom

Correspondence
Kieran M. R. Hunt, Department of Meteorology, University of Reading, Reading, RG6 6BB, United Kingdom.
Email: k.m.r.hunt@reading.ac.uk

Funding information
Newton Fund

Abstract

Low-pressure systems (LPSs) are the primary rainbringers of the South Asian monsoon. Yet, their interactions with the large-scale monsoon circulation, as well as the highly variable land and sea surfaces they pass over, are complex and generally not well understood. In this article, we present a novel, top-down approach to investigate these relationships and quantify their importance in describing LPS behaviour. We also show that, if the approach is sufficiently well posed, it is productive at hypothesis generation. For each of five predictands (i.e., LPS intensification rate, propagation speed/direction, post-landfall survival, peak intensity, and precipitation rate) we train an additive decision-tree ensemble model using the XGBoost algorithm. Shapley value analysis is then applied to the models to determine which variables are important predictors and to establish their relationship with the predictand, with additional analysis following cases of interest. Novel relationships established using this technique include that LPS vorticity intensifies preferentially in the early morning at the same time as the peak in the diurnal cycle of their convection occurs, that vertical wind shear suppresses continued growth of strong LPSs, that large-scale barotropic instability plays an important role in both the inland penetration and peak intensity of LPSs, and that LPS propagation depends on the depth of its vortex with shallower LPSs advected by low-level winds and taller LPSs advected by mid-level winds. We also use this framework to identify and discuss potential new avenues of research for monsoon LPSs.

KEYWORDS

depressions, low-pressure systems, machine learning, monsoon, Shapley values, XGBoost

1 | INTRODUCTION

1.1 | Monsoon low-pressure systems

Low-pressure systems (LPSs) are the predominant source of synoptic-scale variability within the South Asian summer monsoon, bringing the majority of its total (Hunt and Fletcher, 2019) and extreme (Thomas *et al.*, 2021) precipitation (Figure 1a). LPSs vary considerably in frequency, occurring 10–20 times per monsoon season (Sikka, 2006; Vishnu *et al.*, 2020), although they typically spin up over the comparatively warm waters of the Bay of Bengal before making landfall over east India and tracking westward or northwestward, eventually dissipating over the subcontinent where they are bound by the Himalayas to the north and dry surface conditions to the west (Figure 1b). The India Meteorological Department has historically classified monsoon LPSs as

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2023 The Authors. *Quarterly Journal of the Royal Meteorological Society* published by John Wiley & Sons Ltd on behalf of the Royal Meteorological Society.

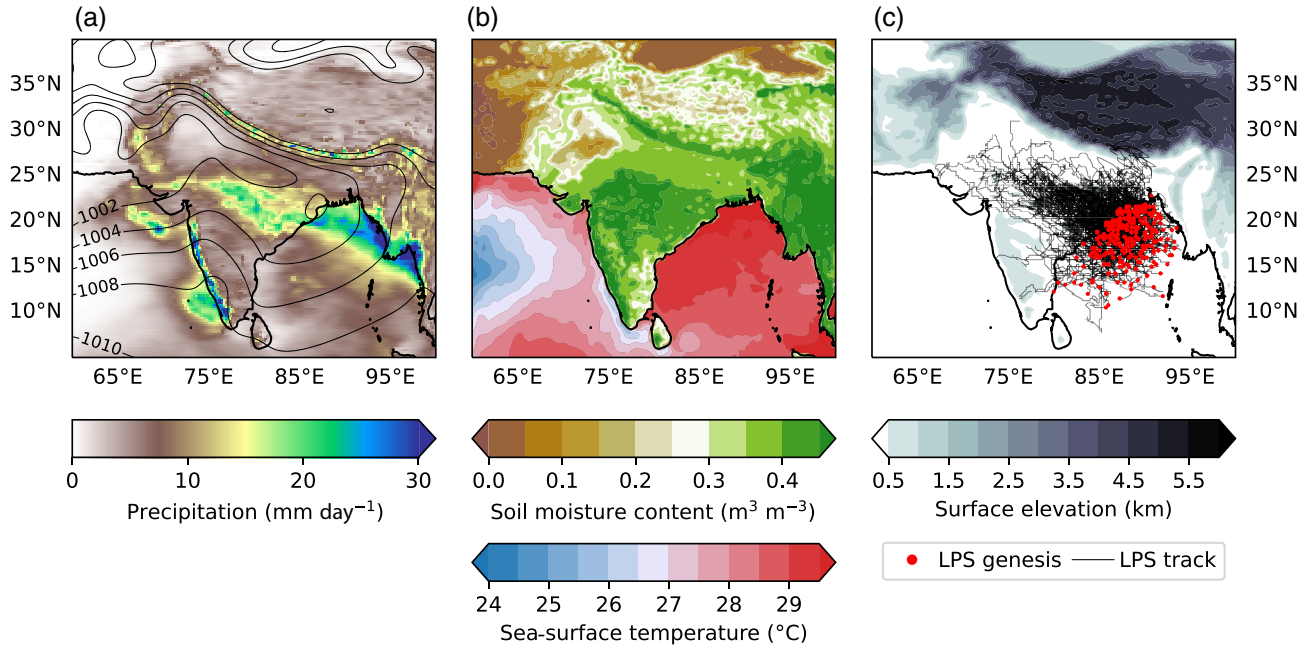

FIGURE 1 Climatological characteristics of the South Asian summer monsoon: (a) mean precipitation and smoothed (using a 1° filter) mean-sea-level pressure; (b) mean soil moisture content between 0 and 7 cm, and mean sea-surface temperature; (c) tracks of the 242 low-pressure systems (LPSs) used in this study, their genesis locations, and the orography of the subcontinent. Computed over June to September 1979–2021 using the fifth-generation European Centre for Medium-Range Weather Forecasts reanalysis.

either low-pressure areas (or simply lows) or monsoon depressions, based on the arbitrary threshold of surface winds exceeding 8.5 m·s^{-1} if over the ocean or two or more closed 2 hPa isobars within 3° of the centre if over land (IMD, 2003). We do not make that distinction in this study since there is no evidence that low-pressure areas and monsoon depressions have fundamentally different responses to environmental stimuli. Instead, where necessary, we present these responses as functions of relative vorticity—a proxy for LPS intensity—giving us a more complete view of the relationship between LPSs and their environment.

Earlier studies that have used LPS tracks to create storm-centred composite fields have revealed a great deal about LPS structure and thermodynamics (Godbole, 1977; Stano *et al.*, 2002; Hurley and Boos, 2015; Hunt *et al.*, 2016a; Hunt *et al.*, 2016b). A warm core aloft is supported by latent heating from both synoptic-scale and convective-scale moist updrafts. This supports a broad, deep, and dense cloud structure as well as heavy precipitation, both of which contribute to a cold core in the boundary layer.

Such a picture is useful but incomplete, as it does not tell us how LPSs interact with their environment; for example, the sea and land surfaces, broader monsoon circulation, and orography. This has led to a number of long-standing open questions, most importantly: Through what processes do LPSs grow and decay, and what explains their westward propagation? The question of LPS intensification has been long debated, with early studies arguing that either barotropic (Goswami *et al.*, 1980; Nitta and Masuda, 1981; Subrahmanyam *et al.*, 1981; Rajamani and Sikdar, 1989) or moist baroclinic (Moorthi and Arakawa, 1985; Salvekar *et al.*, 1986; Krishnakumar *et al.*, 1992) instabilities were responsible. Baroclinic intensification was recently ruled out (Cohen and Boos, 2016), but newer theories have emerged, such as the "moisture-vortex" instability (Adames and Ming, 2018; Adames, 2021), where vortex growth is supported directly by convection. In parallel, a series of idealised model experiments have continued to provide evidence in support of moist barotropic instability, where vortex growth derives from a coupling between moist processes and the meridional shear of the zonal winds (Diaz and Boos, 2019a; Diaz and Boos, 2019b; Diaz and Boos, 2021; Suhas and Boos, 2023). Not only do these theories compete, but they are invariably explained through idealised models, neglecting potentially important factors such as surface fluxes and radiative heating/cooling. In contrast, relatively little research has focused on processes governing LPS decay, despite this being an important control on the time LPSs spend over land, and therefore

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

how much precipitation they contribute to the monsoon. Those studies that have explored LPS decay and inland penetration stress the importance of land surface conditions, either directly through soil moisture (Kishtawal *et al.*, 2013; Hunt and Turner, 2017a) or indirectly through surface flux considerations (Chang *et al.*, 2009).

Similarly, competing theories have emerged about what controls LPS propagation direction and speed. These include downshear vortex stretching through either frictional convergence (Goswami, 1987) or quasi-geostrophic lifting (Sanders, 1984; Chen *et al.*, 2005), interaction with the Himalayas through an image-vortex mechanism (Hunt and Parker, 2016), or beta drift of their midtropospheric potential vorticity (PV) maxima (Boos *et al.*, 2015). The beta drift mechanism is now generally accepted for monsoon depressions, and it has recently been shown that it can be additionally modified by frictional vortex stretching (Hunt and Turner, 2022). However, many weaker LPSs do not possess a secondary PV maximum in the midtroposphere (Deoras *et al.*, 2022), and thus their propagation may not be well explained through advection by mid-level winds.

These are two examples of diagnostic problems (i.e., diagnosing characteristics of an LPS from conditions present at the time), but there are also many prognostic problems, which are usually left for forecast models to address. However, without dedicated experiments, which can be expensive, forecast models do not give us an understanding of the processes involved, which in turn makes diagnosing model biases difficult. In some cases, there are already known issues in the representation of underlying processes in models. One example, as we have already seen, is inland penetration, where land surface processes (in particular, latent heat flux) are important; yet these fluxes are not well represented, even in high-resolution models (Turner *et al.*, 2020). Similarly, forecasts of peak LPS intensity and precipitation are not only affected by the misrepresentation of these land surface processes, but also land–sea interactions (e.g., coastal breezes) and errors arising from convective parametrisation (Mamgain *et al.*, 2018; Podeti *et al.*, 2020).

1.2 | Decision-tree ensemble learning

In our study, we will employ decision tree ensemble models, specifically using a framework known as XGBoost (eXtreme Gradient Boosting; Chen and Guestrin, 2016). Decision-tree models, in their simplest form, are machine learning algorithms that make predictions based on a series of binary questions, the answers leading to further questions, ultimately creating a 'tree' of decisions. Each node of the tree represents a question, with the end points or leaves signifying the final predicted outcomes (Quinlan, 1986; Kotsiantis, 2013; Breiman *et al.*, 2017). These can be used for both classification and regression problems (Loh, 2011).

As trees are binary, they are thus defined only by their depth – the maximum number of nodes a path can pass through in the tree. Most algorithms for training decision trees are greedy (Friedman, 2001), that is, they are built top down, with the algorithm seeking to choose the predictor and threshold that minimises some loss function – typically a root-mean-square error – at each node split. However, to model complex functions without resorting to a single, highly complex decision tree, which is prone to overfitting and instability to minor input variations, we employ an ensemble approach known as "boosting" (Friedman, 2001). The principle of boosting involves the iterative creation of an additive decision-tree ensemble, where each subsequent tree is trained to enhance the collective performance of the previous trees, either up to a defined limit or until the loss reduction saturates (Chen and Guestrin, 2016).

The idea behind this approach is that by aggregating the outputs of multiple "weak" learners (individual trees that perform slightly better than random chance), the model can construct a 'strong' learner that significantly outperforms any of its individual components. The ensemble approach aids in overcoming the problems of overfitting and improving prediction accuracy, which are often associated with single decision-tree models.

XGBoost is an advanced implementation of gradient boosting, a method that constructs the aforementioned ensemble of weak learners in a stage-wise manner. Starting with a simple model, XGBoost iteratively adds new trees that aim to correct the errors made by the existing ensemble. The new trees are fitted to the residual errors of the current ensemble rather than the original targets, a process known as "boosting". By doing so, XGBoost gradually reduces the prediction error, leading to a robust model with extremely strong (often competition-winning) performance metrics (Vidhya, 2016; Dataaspirant, 2020; C-SharpCorner, 2021). While this might appear to be a complex process, a key strength of XGBoost, and decision-tree ensemble models in general, is their interpretability. The decision-making process can be visualised as a series of "if–then" rules, making the model's workings transparent and understandable, even to non-specialists. As we will see, there are various tools to measure the importance of each feature in the model, which provide insights into the factors driving the predictions.

These ensemble models, especially when built using XGBoost, offer a transparent, interpretable model that is robust against overfitting. In the following sections, we will detail the specifics of our approach, explaining how

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

we have built and interpreted our decision-tree ensemble models.

1.3 | Study outline

In this study, we propose a framework of interpretable, or explainable, decision-tree ensemble models to investigate and answer well-posed questions about LPS behaviour using observational data, rather than idealised model experiments. We then demonstrate that framework through five case studies. To ensure a relatively consistent large-scale environment, and because they are the major LPS in terms of their impacts and rainfall contribution over peninsular India (Hunt and Fletcher, 2019), we focus only on LPSs arising over the Bay of Bengal that subsequently make landfall over the Indian subcontinent. The paper is structured as follows. We discuss the data sources used in Section 2. This is followed by an extended methods section (Section 3) in which we discuss how we choose the model inputs (Section 3.1), how we construct our decision-tree ensemble models (Section 3.2), how we tune the models (Section 3.3), how we interpret the results (Section 3.4), and how we make sure these interpretations are safe from cross-correlated inputs (Section 3.5). Our results section is separated into diagnostic models (Section 4.1), in which we investigate models for LPS intensification (Section 4.1.1) and propagation (Section 4.1.2), and forecast models (Section 4.2), in which we investigate models for post-landfall LPS lifetime (Section 4.2.1), peak LPS intensity (Section 4.2.2) and precipitation (Section 4.2.3). Finally, we conclude in Section 5, where we outline key results (Section 5.1) and summarise new research questions revealed by our results (Section 5.2).

2 | DATA

2.1 | ERA5 reanalysis

ERA5 is the fifth generation atmospheric reanalysis of global climate produced by the Copernicus Climate Change Service at the European Centre for Medium-Range Weather Forecasts (Hersbach *et al.*, 2020). Data from ERA5 (available from https://cds.climate.copernicus.eu/cdsapp#!/home) cover the entire globe on a 30 km grid and resolve the atmosphere on 137 levels from the ground up to 80 km in altitude. It covers from January 1950 until the present at hourly frequency. At reduced spatial and temporal resolutions, ERA5 also includes uncertainty information for all variables. We use both single-level and pressure-level variables from ERA5, for both LPS tracking (see Section 2.2) and to compute the predictors and predictands used in training the models (see Section 3.1). For LPS tracking, we use the data at its native hourly frequency, as track fidelity is a high priority. For training and validating the models, we reduce this to six-hourly frequency in order to save disk space.

2.2 | LPS track data

In this study, we apply the LPS tracking algorithm described in (Hunt and Fletcher, 2019) to ERA5 data. We track LPSs by computing the mean relative vorticity in the 900–800 hPa layer, then performing a spectral truncation at T63 to filter out short-wavelength noise. We then identified regions of positive relative vorticity within this field and determined the centroid location for each one. These centroids were then linked in time, subject to constraints in distance (LPS points will not be linked if the implied propagation speed is greater than 100 $km \cdot hr^{-1}$), to form candidate LPS tracks. This algorithm has been used for monsoon LPSs by a number of researchers (e.g., Dong *et al.*, 2020; Martin *et al.*, 2020; Roy and Rao, 2022). These ERA5 LPS track data are available at https://doi.org/10.5281/zenodo.7568990.

For this study, we want a reasonably consistent set of LPS tracks to ensure our research questions remain well-posed. To this end, we include only tracks whose genesis points are over the Bay of Bengal, which terminate over the land of the Indian subcontinent, and which spend a majority of their lifetime between June 1 and September 30. This leaves us with 242 LPSs (Figure 1c) between 1979 and 2021, giving a total of 5,337 observations. These filtered ERA5 LPS track data, including the additional environmental data described in Section 3.1, are available at https://doi.org/10.5281/zenodo.7569057.

3 | METHODS

3.1 | Choice of variables

Table 1 describes the 46 variables used in this study (43 as predictors; six as predictands, given in bold; zonal_speed, merid_speed, and dvo850_dt used as both). The choice of many of these variables is self-explanatory (e.g., we include hour to investigate the importance of the diurnal cycle), but we now include a brief explanation for those that are not. Throughout, those prefixed with "mean_" are taken as an average within 400 km of the LPS centre, and those prefixed with "mcz_" are averaged over (75–85° E, 18.5–27° N). The rationale behind the 400 km radius lies in mitigating uncertainties linked to LPS track location,

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Definitions of variables used in this study.

Variable name	Description	Units
year	Time step year	
month	Time step month	
hour	Time step hour of day	UTC
x	Longitude of LPS centre	° E
y	Latitude of LPS centre	° N
over_land	Flag for LPS centre	Boolean
orography_height	Elevation of land surface under centre	m
acc_land_time	Accumulated time where over_land = True	hr
total_land_time	Final value of acc_land_time for a given LPS	hr
mean_land_frac	Fraction of area within 400 km that is over land	
mean_skt	Surface temperature	K
mean_land_skt	Land surface temperature (NaN over ocean)	K
mean_sst	Sea surface temperature (NaN over land)	K
mean_swvl1	Soil moisture in the top layer (<7 cm; NaN over ocean)	$m^3 \cdot m^{-3}$
mean_swvl2	Soil moisture in the second layer (7–28 cm; NaN over ocean)	$m^3 \cdot m^{-3}$
mean_swvl1_grad	Mean absolute gradient of swvl1 ($\lvert\nabla$swvl1$\rvert$)	$m^3 \cdot m^{-4}$
mean_swvl2_grad	Mean absolute gradient of swvl2	$m^3 \cdot m^{-4}$
mean_u850	850 hPa zonal wind	$m \cdot s^{-1}$
mean_u500	500 hPa zonal wind	$m \cdot s^{-1}$
mean_u200	200 hPa zonal wind	$m \cdot s^{-1}$
mean_v850	850 hPa meridional wind	$m \cdot s^{-1}$
mean_v500	500 hPa meridional wind	$m \cdot s^{-1}$
mean_v200	200 hPa meridional wind	$m \cdot s^{-1}$
zonal_speed	x-component of LPS propagation vector	$km \cdot hr^{-1}$
meridional_speed	y-component of LPS propagation vector	$km \cdot hr^{-1}$
mcz_tcwv	Mean total column water vapour over monsoon trough	$kg \cdot m^{-2}$
mean_q_850	850 hPa specific humidity	$m^3 \cdot m^{-3}$
mean_cape	CAPE	$J \cdot kg^{-1}$
mcz_cape	Mean CAPE over the monsoon trough	$J \cdot kg^{-1}$
mean_dthetae_dp_900_750	d(theta_e)/dp between 900 and 750 hPa	$K \cdot hPa^{-1}$
mean_dthetae_dp_750_500	d(theta_e)/dp between 750 and 500 hPa	$K \cdot hPa^{-1}$
olr_90	90th percentile of negative OLR (i.e., ~90th percentile of cloud top height)	$W \cdot m^{-2}$
olr_75	75th percentile of negative OLR within 400 km	$W \cdot m^{-2}$
olr_50	50th percentile of negative OLR within 400 km	$W \cdot m^{-2}$
mean_prcp_400	Mean precipitation within 400 km over the next 6 hr	$mm \cdot hr^{-1}$
qshear_850	Meridional shear of 850 hPa specific humidity over India	$m^3 \cdot m^{-3} \cdot deg^{-1}$
ushear_850	Meridional shear of 850 hPa zonal wind over India	$m \cdot s^{-1} \cdot deg^{-1}$
qshear_850_background	qshear_850 averaged over the previous 10 days	$m^3 \cdot m^{-3} \cdot deg^{-1}$
ushear_850_background	ushear_850 averaged over the previous 10 days	$m \cdot s^{-1} \cdot deg^{-1}$

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 (Continued)

Variable name	Description	Units
mean_vort_850	Relative vorticity in 850 hPa layer	10^{-5} s^{-1}
mean_vort_700	Relative vorticity in 700 hPa layer	10^{-5} s^{-1}
mean_vort_500	Relative vorticity in 500 hPa layer	10^{-5} s^{-1}
vortex_depth	(mean_vort_500 × mean_vort_700)/(mean_vort_850)^2	
dvo850_dt	Rate of change of mean_vort_850	10^{-5} s^{-1} $\cdot day^{-1}$
reached_peak	False if peak_vorticity has not been reached yet, else True	Boolean
peak_vorticity	Largest value of mean_vort_850 attained by a given LPS	10^{-5} s^{-1}

Note: All variables prefixed with "mean_" are taken as an average within 400 km of the LPS centre. mcz_cape and mcz_tcwv are averaged over the monsoon core zone (75–85° E, 18.5–27° N). qshear_850, ushear_850, and their backgrounds are averaged over 5° longitude either side of the LPS centre, with the gradient computed between 10° N and 27° N. Variables marked in bold are predictands.
Abbreviations: CAPE, convective available potential energy; LPS, low-pressure system; NaN, not a number; OLR, outgoing long-wave radiation.

and in accounting for the spatial heterogeneity of various variables across the LPS. For instance, variables such as precipitation and outgoing long-wave radiation (OLR) demonstrate significant spatial variability and their most extreme values may not necessarily be located at the LPS centre, but rather towards its southwest. The choice of a 400 km radius thus provides a representative spatial scale that captures most of the variable variance within the LPS, smoothing out the extreme values, and thereby offering a more balanced picture of the LPS dynamics (Hurley and Boos, 2015; Hunt *et al.*, 2016a).

The soil moisture gradient terms (mean_swvl1_grad and mean_swvl2_grad) are included following work by Barton *et al.* (2020), which suggested that regions of high soil moisture gradient support low-level convergence within the monsoon, and hence are associated with convective initiation—which may in turn support LPS growth and precipitation. In a similar vein, we also include several terms that measure moist thermodynamic instability within the LPS itself (mean_cape, mean_dthetae_dp_900_750, mean_dthetae_dp_750_500, mean_q_850), as well as the monsoon in general (mcz_tcwv and mcz_cape). The use of convective available potential energy (CAPE) from a reanalysis must be justified, since tropical lapse rates in models are prone to biases (Gillett *et al.*, 2000). However, recent studies have found that ERA5 captures CAPE well except for extreme values (Taszarek *et al.*, 2021; Wang *et al.*, 2021). As we average CAPE over large regions, the potential underestimation of the right tail is unlikely to be problematic. Therefore, for the sake of dataset consistency, we use ERA5 CAPE.

These variables only quantify the potential for convective initiation and growth, so we also include several OLR terms (olr_90, olr_75, olr_50) that quantify the depth and spread of established convection, as these may also be useful for models predicting precipitation and whether the LPS is near its peak intensity. The use of precipitation from a reanalysis must also be justified, since earlier reanalyses have substantial biases in tropical precipitation (Bosilovich *et al.*, 2008; Ma *et al.*, 2009). Fortunately, these are much improved in ERA5 to the point where the errors are, in some regions, comparable to those from satellite estimates (Xu *et al.*, 2022). Given we also average precipitation over large areas, we are confident in using ERA5 precipitation for our analysis. This gives the additional advantage of a long and continuous dataset. However, for the interested reader, the dataset released with this article (see Section 2.2) also includes Integrated Multi-satellite Retrievals for Global Precipitation Measurement satellite precipitation.

Finally, we include four terms that measure the large-scale meridional shear in lower-tropospheric zonal wind and specific humidity (qshear_850, ushear_850) and their respective background values (qshear_850_background, ushear_850_background). This follows Suhas and Boos (2023), who used these terms to describe the instability available for moist barotropic growth (Diaz and Boos, 2019b) and moisture-vortex instability (Adames and Ming, 2018). As we will show in Section 3.5, not all variables are used or needed in all decision-tree models.

3.2 | Setting up and running the decision-tree ensemble

As the five predictands in this study are continuous variables, we utilise regression trees, which are designed to predict continuous targets, as opposed to classification trees that predict discrete targets (such as labels). For a full discussion of the differences, the reader should follow

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

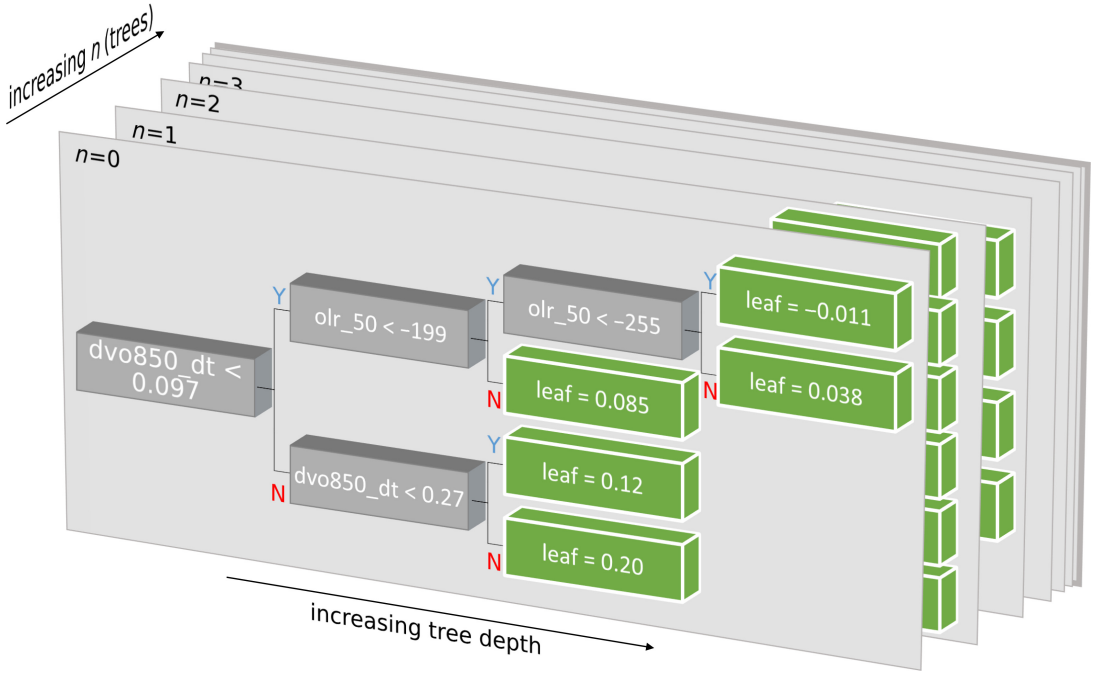

FIGURE 2 Schematic demonstrating the structure of an example additive decision-tree ensemble, trained by XGBoost to predict mean_prcp_400, but with maximum tree depth limited to 3 for clarity. Grey boxes represent the binary decisions being made based on meteorological thresholds; for example, an olr_50 below $-199\ \mathrm{W{\cdot}m^{-2}}$. The green boxes, representing the leaves of the tree (after which no further decisions can be made along that pathway), contain the estimated value of mean_prcp_400. The predicted mean_prcp_400 for a given time step is then computed by parsing each decision tree and summing the n obtained leaf values.

Loh (2011). Figure 2illustrates an example of a simple additive decision-tree ensemble created by our implementation of XGBoost.

During model training, we impose two commonly applied constraints. First, we use an 80–20 train–test split, where 20% of randomly chosen data is withheld from the training process for validation, thus preventing overfitting. Second, an early stopping criterion is applied, where if the training root-mean-square error (RMSE) does not improve after the addition of 10 more trees then the model is considered to have reached its optimal complexity and training is terminated.

3.3 | Hyperparameters and Bayesian optimisation

As with many machine-learning algorithms, XGBoost is controlled by a set of hyperparameters that affect the learning rate and complexity of the model. We have already discussed two: maximum tree depth and total number of trees in the ensemble. Three more are of interest: the learning rate, proportional to the step size used by the gradient descent solver; and two regularisation terms, one penalising the addition of unnecessary nodes (γ) and one penalising both highly asymmetric nodes (i.e., splitting off only a handful of cases) and the addition of new predictors to the model (α).

A grid search over five hyperparameters is slow, so we turn to a Bayesian approach to seek an optimal tuning configuration (Mockus, 1994; Jones *et al.*, 1998; Brochu *et al.*, 2010; Wang *et al.*, 2020). Bayesian optimisation first reduces the problem to the minimisation of a single multivariate objective function, $\mathrm{RMSE} = F(d, n, \epsilon, \alpha, \gamma)$, where d is the maximum tree depth, n is the number of trees in the model, and ϵ is the learning rate. It then selectively samples F by iteratively computing new sampling points using so-called "acquisition functions" which compromise between regions where F is already known to be small and regions where there is low certainty in the predicted value of F. In this study, we use the Python implementation described at http://krasserm.github.io/2018/03/21/bayesian-optimization/. There are several more hyperparameters associated with XGBoost training, but results are not typically sensitive to these except in edge cases (Wang and Chen, 2019). For these hyperparameters, we retain the Python implementation defaults (https://xgboost.readthedocs.io/en/stable/parameter.html [1]).

[1] Archived at https://web.archive.org/web/20230124160343/https://xgboost.readthedocs.io/en/stable/parameter.html.

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Iterative improvements to the total_land_time model root-mean-square error (RMSE) using Bayesian hyperparameter tuning.

Iteration	RMSE	α	γ	Learning rate	Tree depth	n (trees)
1	40.02	13.2	3.95	0.029	4	89
2	36.14	14.3	3.37	0.61	3	157
3	34.36	16.2	0.69	0.17	4	143
6	32.53	17.1	2.83	0.14	5	180
24	31.91	9.25	0.010	0.42	7	91
34	31.37	2.55	0.045	0.059	7	145
54	30.63	9.41	4.75	0.27	6	127
97	30.56	17.4	1.76	0.14	6	102
99	30.53	9.17	4.79	0.27	6	127

Note: α is an L1-regularisation term (higher values penalise the model for adding new predictors) and γ is the minimum loss reduction required to make a further partition on a leaf node. Higher values of α and γ correspond to more conservative training but a reduced vulnerability to overfitting. Learning rate is proportional to the step size used in the solver when iterating weights during learning. Lower values of learning rate mean the model converges more slowly during training but is more likely to converge on a global minimum. The target is minimisation of RMSE, which is computed here using a fivefold cross-validation. The Bayesian optimisation algorithm was run for 100 iterations, but we show only those iterations that reduce RMSE.

We must also choose the search domains for each of the hyperparameters. The lower bounds are trivially 0 or 1, as α, γ, and ϵ are positive definite and, by definition, trees must have a depth of at least 1 and ensemble models must have at least one member. The upper bounds are not trivial, so are either guided by computational cost (fixing an upper limit of 1,000 trees, with a maximum depth of 10) or values where model performance is significantly degraded (fixing an upper limit of $\alpha = 20$, $\gamma = 20$, $\epsilon = 1$). Each iteration of the Bayesian optimisation is then run with fivefold cross-validation to ensure robust computation of the objective function. Table 2 shows example improvements in tuning over 100 iterations for the total_land_time model.

3.4 | Shapley values and interpretability

To make our decision-tree ensemble models interpretable, and hence explainable, we use Shapley value theory (Shapley, 1953; Roth, 1988; Lundberg and Lee, 2017). Originating in cooperative game theory, Shapley values were originally devised as a way to assign payouts to players depending on their contribution towards the total payout by considering how different permutations of players perturb the outcome. In black- or grey-box model,s such as our decision-tree ensemble, Shapley values estimate the marginal contribution from a given predictor in forcing a prediction away from the distribution mean. The sum of all the Shapley values for all predictors for a given prediction $\hat{Y}$, therefore, is equal to the difference between the predicted value and the predictand mean; that is, $\hat{Y} - E(Y)$.

In this study, we use the "shap" Python package (https://github.com/slundberg/shap), which contains TreeSHAP (Lundberg *et al.*, 2020), a highly optimised algorithm for computing Shapley values for decision-tree ensembles.

3.5 | Pruning redundant variables (feature selection)

When training our models, we must first take care to remove, or prune, highly correlated predictors. This is important for interpretability: as the correlation between two predictors increases, they become increasingly degenerate (i.e., interchangeable) within the model. Shapley values are additive, and are therefore shared between highly interchangeable predictors. This can be misleading (leading to underestimates) when it comes to interpreting their impact on model predictions. We use two methods to mitigate this problem. Furthermore, careful pruning has been shown to drastically improve model performance (Sorscher *et al.*, 2022).

First, we compute the linear correlation coefficient between all variables (for which a large subset is given in Figure 3). We identify variables in pairs or groups that share at least 50% of explained variance (i.e., $r^2 > 0.5$). Variables within these groups are removed as predictors in all models until only one per group remains. Using this method, we removed six variables: vo_700, owing to high correlation with vo_850 and higher correlation with vo_500 than vo_850; olr_70, due to high correlation with olr_90 and olr_50; qshear_850 and ushear_850, owing to their high correlations with qshear_850_background and ushear_850_background respectively, choosing to retain the background terms because they better describe

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

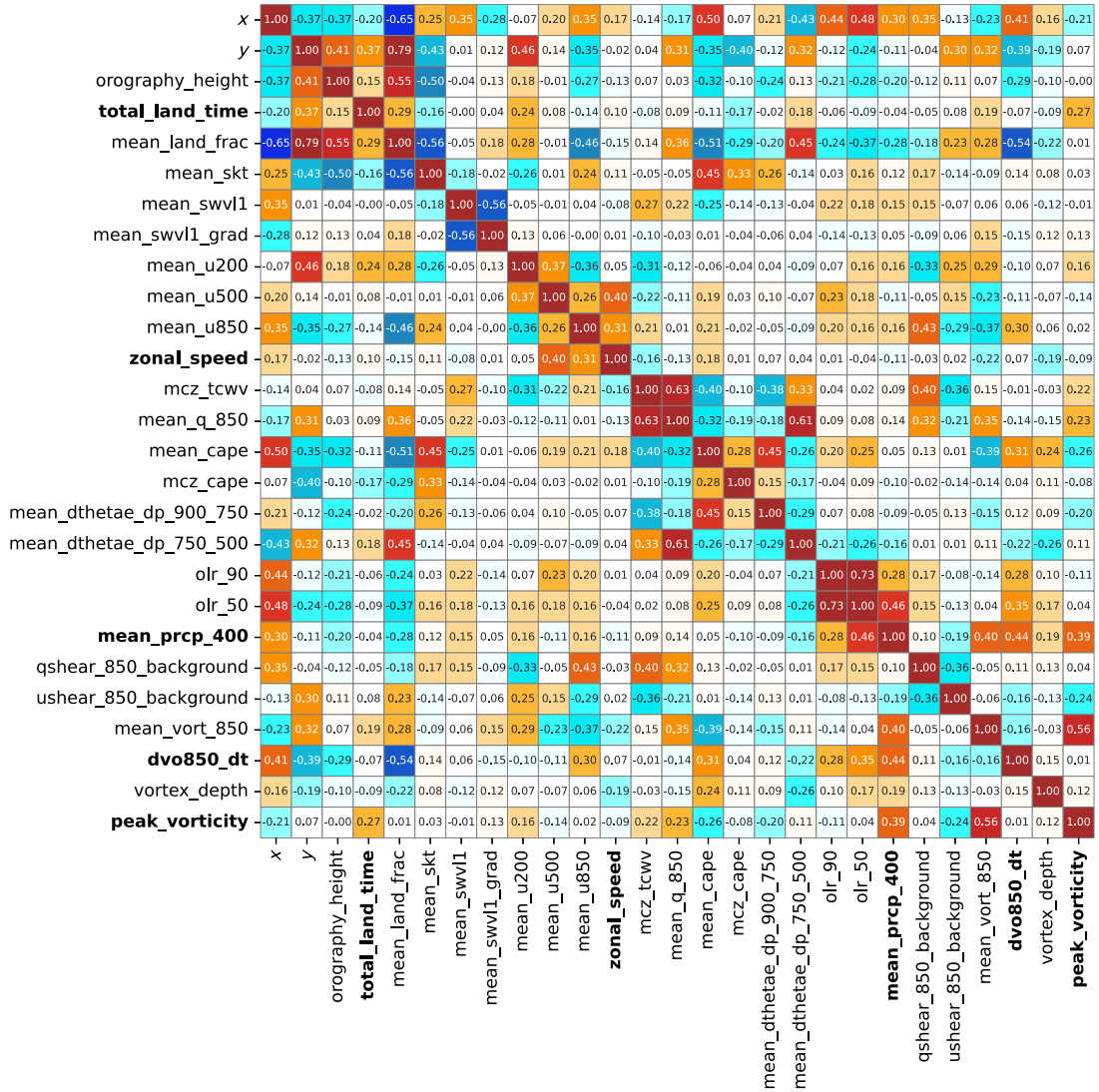

FIGURE 3 Correlation coefficients between selected predictor and predictand variables, loosely grouped by type (geographical, circulation, thermodynamical, vortical). Predictand variables are labelled in bold. Definitions of all variables are given in Table 1.

the underlying processes we want to include; and acc_land_time and over_land, owing to their high correlation with *x*, *y*, and mean_land_frac, which are clearer descriptors of the underlying geography. We retain mean_land_frac (despite its high correlation with *y*) and olr_90 and olr_50 (despite their high correlation with each other) because these pairs describe fundamentally different properties about the land surface and convection respectively, which may be important to distinguish between later. This leaves us with a maximum of 37 predictors for each model.

Second, we use hierarchical clustering to prune redundant predictors on a predictand-by-predictand basis. For each predictand, the process is as follows. We start by training a full model (37 predictors), computing and storing the respective sets of Shapley values for each predictor. Then, we train 37 univariate models, one for each predictor. Each of these models is then run again 36 times, with the original predictor sequentially replaced by the other predictors. If the replaced-predictor model explains at least 50% of the variance of the original-predictor model, we consider the pair of predictors redundant and remove the one with the lowest mean absolute Shapley value from the initial full model run. All predictors must be normalised by their standard deviations beforehand. We obtain very similar results for all six predictands, removing year due to similarities with mean_skt and mean_q_850; removing either mean_swvl1 at the expense of mean_swvl2 or vice versa (and the same with their gradients), and removing zonal_speed and merid_speed owing to respective similarities with mean_u850/mean_u500 and mean_v850/mean_v500.

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4 | RESULTS

In this section we discuss the performance, predictor importance, and implications of each of the five models in turn. The models are grouped into diagnostic (dvo850_dt and zonal_speed/meridional_speed) and forecast models (total_land_time, peak_vorticity, and mean_prcp_400). Diagnostic models are trained to predict LPS characteristics simultaneous with the predictors, whereas forecast models are trained to predict some LPS characteristic ahead of time.

4.1 | Diagnostic models

4.1.1 | Growth and decay

We start with a brief overview on how LPSs behave over their lifetime (Figure 4). The intensification rate starts high, falling throughout the LPS lifetime, leading to an intensity maximum a little after halfway between their genesis and lysis. This intensity peak coincides with the period of heaviest LPS rainfall and occurs at about the same time that most members of the LPS composite make landfall.

The first decision-tree ensemble model is trained to predict LPS intensification and weakening, which we measure through the rate of change of low-level relative vorticity (dvo850_dt). Figure 5a shows that the model generally performs well over the training data, except for underestimating cases of rapid intensification (observed dvo850_dt $\gtrsim 2 \times 10^{-5}\,s^{-1}{\cdot}day^{-1}$). The distributions of Shapley values for each predictor are shown in Figure 5b, sorted by their relative impact (i.e., the mean magnitude of their Shapley values). For clarity, we only show the 12 predictors with the highest impact. We also colour each Shapley value distribution according to the standardised value of the underlying predictor variable.

Several important forcings emerge. First, the location of the LPS is clearly important: systems over land or near orography tend to weaken (i.e., low mean_land_frac, high latitude (y), low longitude (x), and high orography_height),

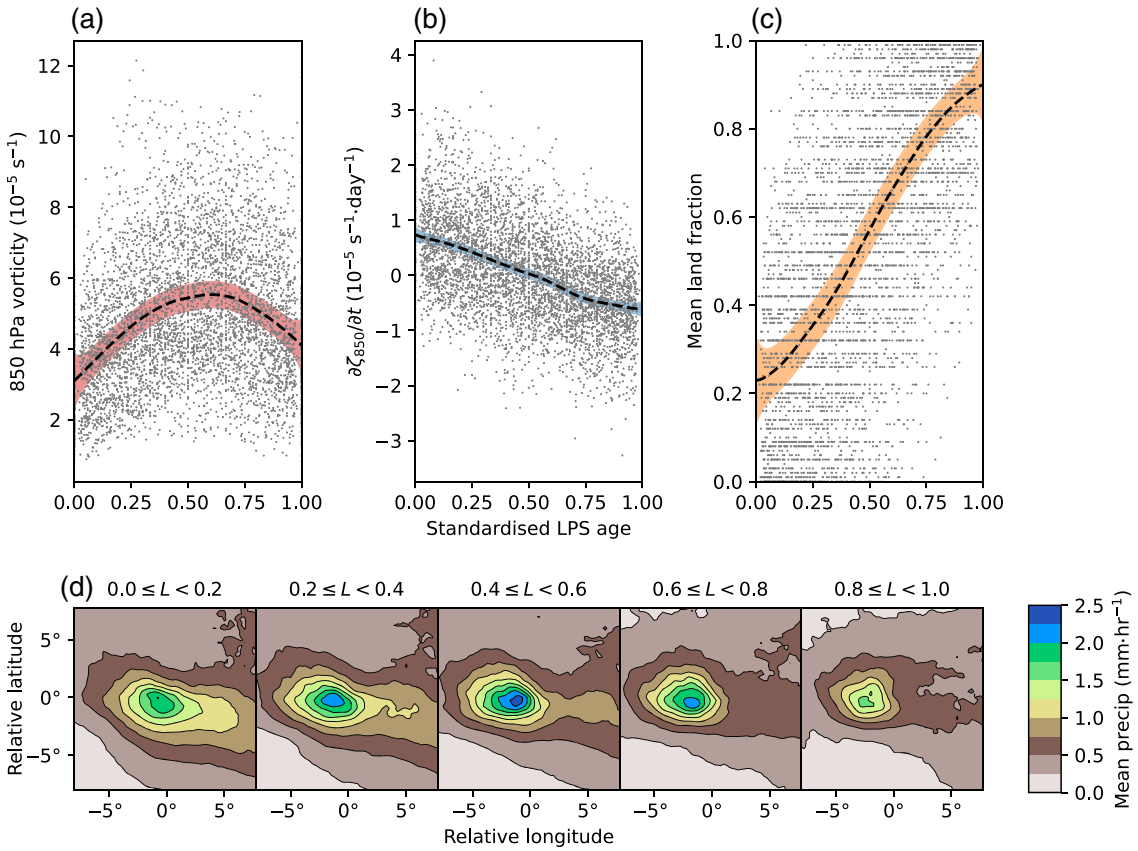

FIGURE 4 Selected features of monsoon low-pressure systems (LPSs) as a function of their standardised age (i.e., genesis at 0, lysis at 1): (a) 850 hPa relative vorticity, (b) the rate of change of 850 hPa relative vorticity, and (c) the land fraction, all computed as means within 400 km of the LPS centre. Each grey dot represents a single LPS time point, with a lowess smoothing given by a dashed black line. The coloured bands represent the 99% confidence intervals of the smoothing lines. (d) Mean LPS-centred precipitation as a function of standardised age pentile.

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

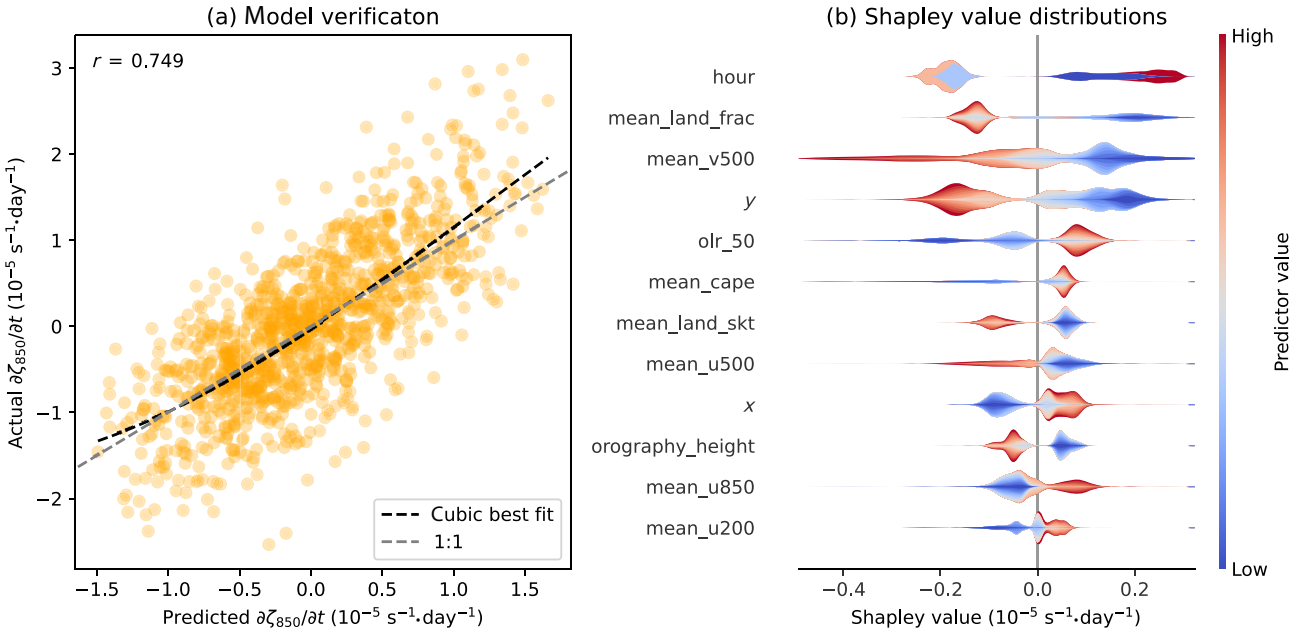

FIGURE 5 Verification and interpretation of the decision-tree model predicting dvo850_dt (the rate of change of mean 850 hPa relative vorticity within 400 km of the low-pressure system centre). (a) Model predictions are plotted against observed values, with the grey dashed line denoting a 1:1 relationship and the black dashed line showing a cubic best fit. The linear regression coefficient between the actual and predicted values is given in the top left. (b) Interpretation of the relative importance of predictors in the model is shown through their Shapley value distributions. The predictor variables are sorted by the mean of their absolute Shapley values, with the distributions coloured according to the underlying value of the variable. This model converged after 64 rounds.

whereas those over the ocean tend to intensify. This is also reflected in the importance of mean_v500: high values push LPSs over land and towards the Himalayan orography, resulting in their weakening. Second, active convection (large mean_cape and olr_50) supports LPS intensification, presumably through increased latent heat release in the midtroposphere. Neither of these results are surprising or novel, but they show the feasibility of using explainable decision trees to understand process drivers. The third important forcing—the diurnal cycle (represented by hour)—is more surprising and, therefore, merits further exploration.

In Figure 6 we plot the diurnal cycle of dvo850_dt for all LPSs as a function of location, separating them into land (at least 75% of surface within 400 km is land), ocean (at least 75% of surface within 400 km is ocean), and coast (otherwise). LPSs in all three locations show a maximum in intensification at about 0600 h local time. LPSs over land have a another, slightly stronger maximum at about 1800 h local time. The early morning peak is also present in the diurnal cycle of both tropical cyclone (Bowman and Fowler, 2015) and monsoon depression (Hunt *et al.*, 2016b) precipitation, where it is attributed to increased instability caused by radiative cooling of upper level clouds during night-time.

For LPSs over the ocean, this peak in the growth rate of lower-tropospheric vorticity is aligned with the climatological peak in diurnal convection and precipitation over tropical oceans (Yang and Slingo, 2001; Liu and Zipser, 2008). For LPSs over land, when convection typically occurs in the late afternoon, these peaks are not aligned. This means that LPSs over land do not benefit from a potential positive feedback between the two processes (diurnal heating of the surface and radiative cooling aloft), which may contribute to their weakening post-landfall by damping deep convection near the centre. Further investigation of this relationship is a topic we leave for future research.

Finally, we note that the predictors representing large-scale barotropic instability, ushear_850_background, and moisture-vortex instability, qshear_850_background, are not considered important by the model. This may be because they are related to the development of, and therefore partially correlated with, other features important to LPS growth and decay, such as OLR. However, as these variables were not sifted out during our preliminary correlation analysis, nor during the subsequent hierarchical clustering, we can be fairly sure this is a robust result. The implication, therefore, is that idealised models may not be sufficiently complete tools to diagnose modes of LPS intensification in the context of their complex environment.

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

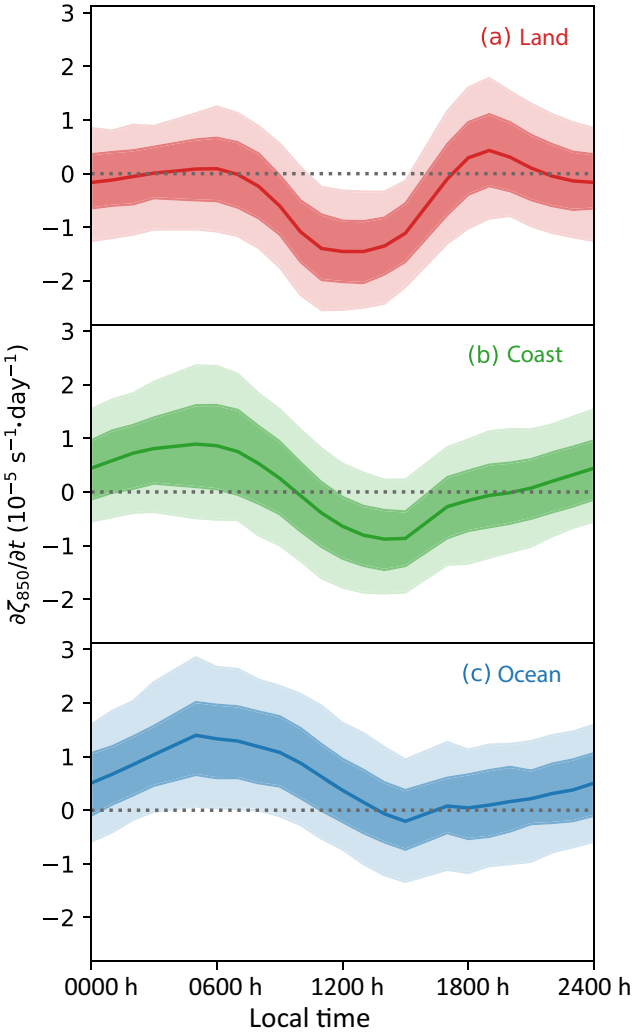

FIGURE 6 Mean diurnal cycle of low-pressure system (LPS) $\partial\zeta_{850}/\partial t$ (the rate of change of 850 hPa relative vorticity averaged within 400 km of the LPS centre). Partitioned according to LPS location: (a) land—at least 75% of the surface within 400 km of the centre is land; (b) coast—between 25% and 75% of the surface within 400 km of the centre is land; (c) less than 25% of the surface within 400 km of the centre is land. Inner and outer filled bands indicate the interquartile and 10th–90th percentile ranges respectively. Data for 0900 UTC and 2100 UTC (0300 h and 1500 h local time respectively) have been linearly interpolated from neighbouring points to account for discontinuities introduced by the data assimilation window edges in the fifth generation European Centre for Medium-Range Weather Forecasts reanalysis.

4.1.2 | Propagation speed and heading

We now move on to our second and third diagnostic models, which we treat together: predicting the zonal and meridional components of LPS propagation vectors. These models do not perform as well as that for dvo850_dt, with correlation coefficients of 0.59 and 0.57 between predicted and observed speeds respectively (Figure 7). Even so, they perform well enough to support an initial investigation into the drivers of LPS propagation speed.

Both models select steering winds at 850 hPa and 500 hPa as the two most important predictors, with both assigning roughly equal importance to each. However, if we were to take the results of Boos *et al.* (2015) at face value, we would expect a much stronger contribution from the 500 hPa winds than the 850 hPa winds. The resolution lies in the secondary contributions from vortex_depth, the ratio of mid-level to low-level vorticity, and x, the longitude. To show why, we separate LPSs into the top and bottom quartiles of (a) mean_u500 and (b) vortex_depth, bin them onto a 1.5° × 1.5° grid, and then plot their mean propagation vectors as a function of location (Figure 8). Despite the fact that mean_u500 and vortex_depth are uncorrelated (see Figure 3), LPSs in the tails of their distributions behave remarkably similarly. In other words, LPSs with a small vortex depth (blue arrows)—that is, those with relatively weak vorticity at 500 hPa—propagate very similarly to LPSs present when the mid-level easterlies are weak (red arrows), even though the two populations do not significantly overlap. Therefore, we hypothesise that the propagation theory developed by Boos *et al.* (2015)—that LPSs (specifically depressions, which make up roughly the strongest quartile of LPSs) propagate through advection of their midtropospheric PV maximum by 500 hPa winds and beta drift—only holds for LPSs with substantial vortex depth; that is, LPSs whose 500 hPa relative vorticity is not substantially smaller than their lower-tropospheric maximum.

We can test this hypothesis directly using the Shapley values from the zonal_speed model (Figure 9). We show, for reference, the storm-centred composite vertical structure of PV for the LPSs used in this study (Figure 9a). This structure is morphologically very similar (showing a midtropospheric maximum about 500 hPa and lower-tropospheric maximum at about 750 hPa) to that presented for stronger monsoon LPSs and monsoon depressions in Hurley and Boos *et al.* (2015) and Hunt *et al.* (2016a) respectively, but the magnitude is about 20% weaker. The difference in PV structure between LPSs with large and small vortex depth is associated almost entirely with the 500 hPa maximum. This is in contrast to the difference between LPSs in the top and bottom quartiles of 500 hPa (not shown), where the signal extends into the lower troposphere, and thus demonstrates that vortex depth is a useful variable to isolate the effects of the midtropospheric PV maximum. We leave the follow-up question regarding why some LPSs develop a strong 500 hPa maximum but others do not for future research.

We leverage this in Figure 9c, which shows the gridwise correlation coefficients of PV with the Shapley values of both vortex_depth and mean_vort_850. Recall

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

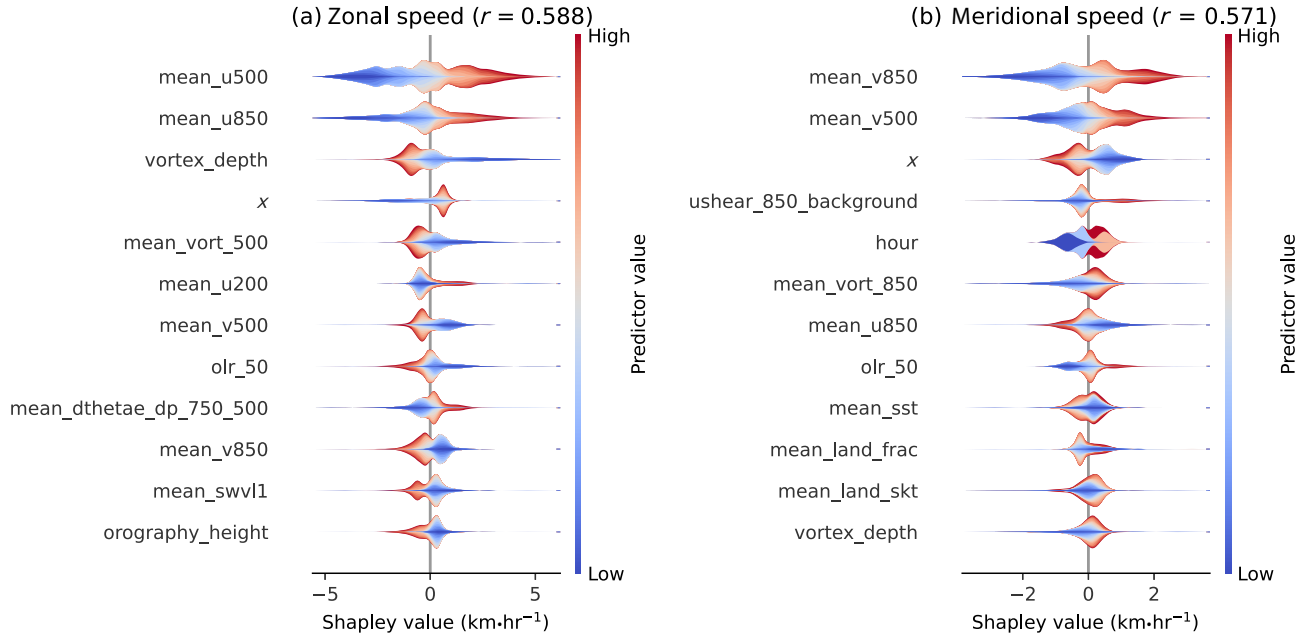

FIGURE 7 Interpretation of the relative importance of predictors in the (a) zonal_speed (westward component of low-pressure system propagation vector) and (b) meridional_speed (northward component of low-pressure system propagation vector) models, shown using their respective Shapley value distributions. Linear correlation coefficients between the actual and predicted values are given in the subfigure titles. The two models converged after 62 and 148 rounds respectively.

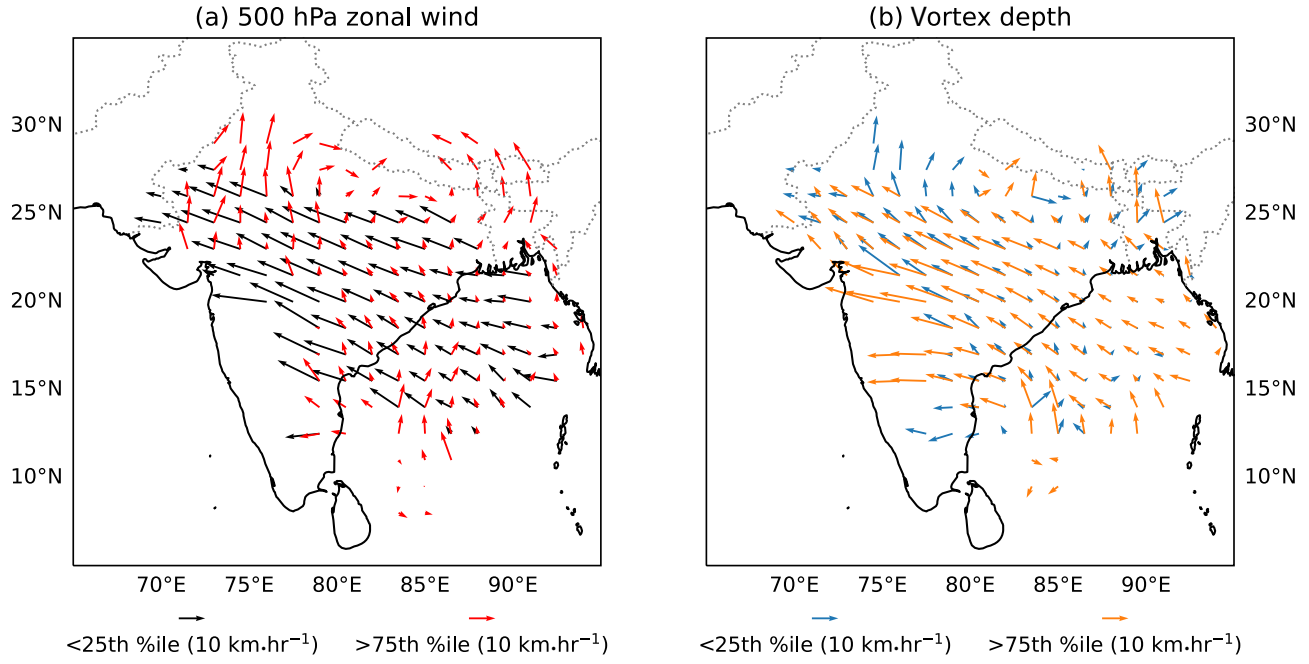

FIGURE 8 Mean low-pressure system (LPS) propagation as a function of (a) mean 500 hPa zonal wind within 400 km of the LPS centre and (b) LPS vortex depth. LPSs in the top and bottom quartile of each variable are binned onto a 1.5° × 1.5° grid. The mean propagation vector for LPSs at each grid point is then plotted, so long as at least five LPSs are present.

that as the majority of LPSs propagate westwards they have a negative zonal_speed, and so predictors that increase zonal propagation speed will have negative Shapley values (and hence negative correlation coefficients). The vortex_depth Shapley values are significantly correlated with midtropospheric PV (peaking at about 450 hPa), whereas mean_vort_850 Shapley values are significantly correlated with lower-tropospheric PV

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

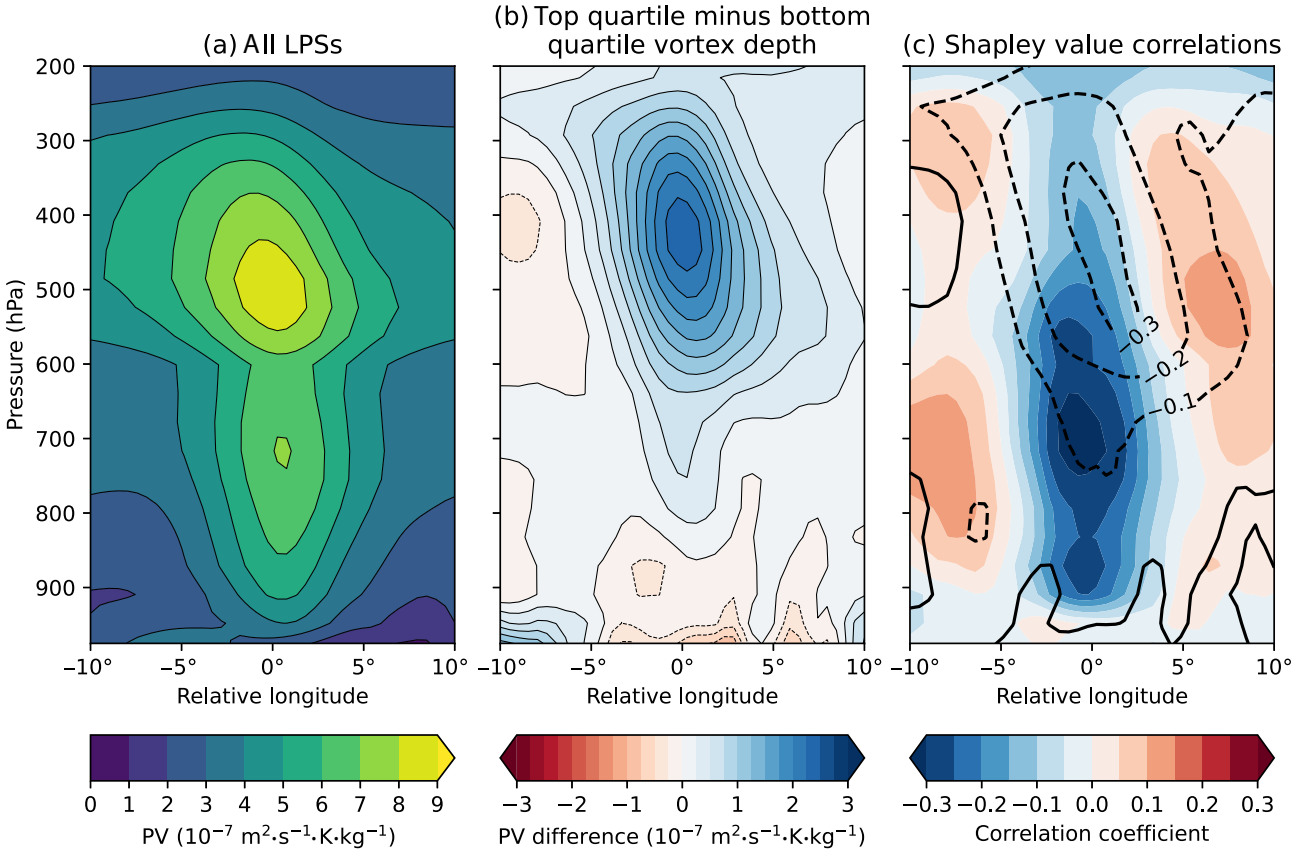

FIGURE 9 The relationship between the vertical structure of low-pressure system (LPS) potential vorticity (PV) and zonal propagation speed. (a) Zonal cross-section of storm-centered PV for all LPSs; (b) difference in composite PV between LPSs in the top and bottom quartiles of the vortex depth distribution; (c) correlation coefficient between PV and the Shapley values for mean_vort_850 (filled contours) and vortex_depth (line contours) in the zonal_speed model.

(peaking at about 700 hPa). This supports our hypothesis that LPS propagation is driven by advection of the midtropospheric PV maximum only if that maximum is not small compared with the lower-tropospheric maximum; otherwise, propagation is more likely to be driven by advection of the lower-tropospheric maximum.

This discussion covers many of the predictor terms in Figure 7, but there are a few others that merit attention. In particular, the relatively high impact of olr_50 (with deeper convection linked to faster propagation) hints at the secondary role played by off-centre vortex stretching in supporting LPS movement. This was linked to monsoon intraseasonal variability by Hunt and Turner (2022), but has also previously been mooted as the primary propagation mechanism (Sanders, 1984; Chen *et al.*, 2005). In the zonal_speed model, the high impacts of mean_dthetae_dp_750_500 (mid-level instability) and mean_swvl1 (local moisture source) suggest that such convection is supported by large-scale conditions. We will discuss this further when we come to analyse the precipitation model later.

Finally, we note the caveat that the two propagation models have the weakest performance among all models presented in this article, each explaining about one-third of the variance in propagation speed. More work is needed to assess whether this is due to a particular variable (or set of variables) missing from the predictors (e.g., background wind speeds—as we have used for the shear variables), or whether existing predictors should be used at different pressure levels.

4.2 | Forecast models

4.2.1 | Inland penetration

The first of our predictive models is trained to predict the total amount of time an LPS will spend over land. As with the diagnostic models in the previous section, this model is trained on single time points, meaning that multiple predictions are made for a given LPS (analogous to forecasts at different lead times). The model performs well,

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

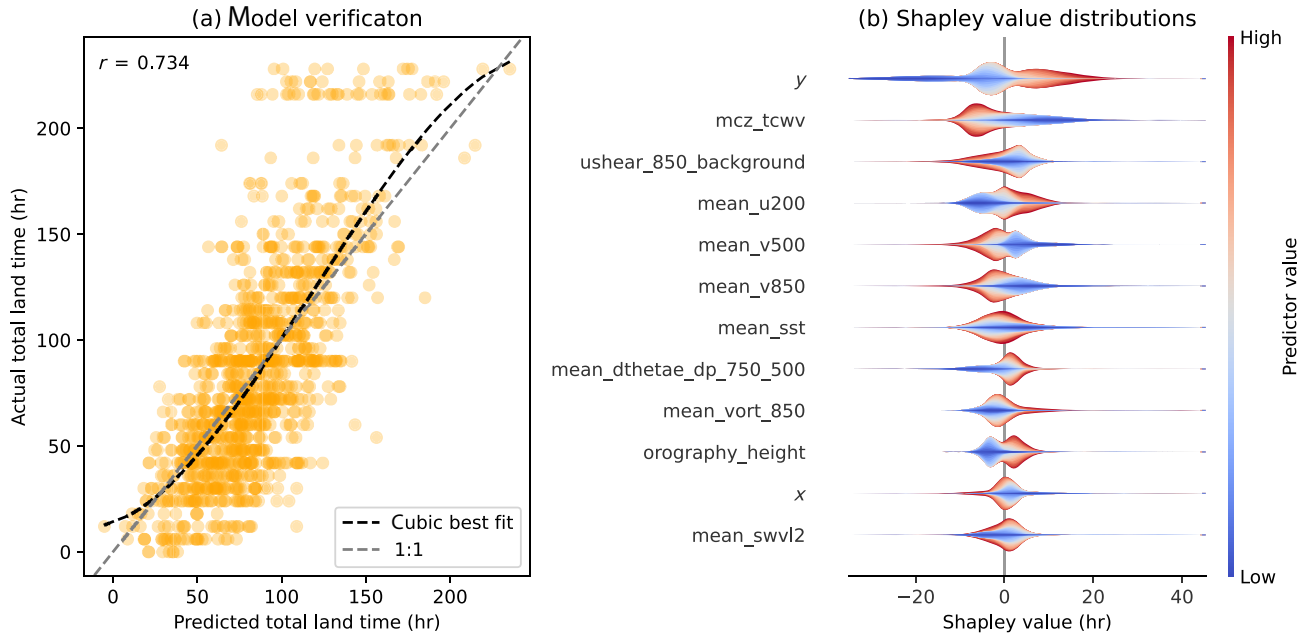

FIGURE 10 Verification and interpretation of the decision-tree model predicting total_land_time (the total time spent over land by a low-pressure system). (a) Model predictions are plotted against observed values, with the grey dashed line denoting a 1:1 relationship and the black dashed line showing a cubic best fit. The linear correlation coefficient between the actual and predicted values is given in the top left. (b) Interpretation of the relative importance of predictors in the model is shown through their Shapley value distributions. The predictor variables are sorted by the mean of their absolute Shapley values, with the distributions coloured according to the underlying value of the variable. This model converged after 317 rounds.

explaining over half the variance (Figure10a), although it tends to underestimate large values of total_land_time. The most important predictors, sorted by their Shapley values, are given in Figure 10b. Many of these entries are quite intuitive. High values of *y* support longer total_land_time because there is more land at higher latitudes and because LPSs typically have a small northward component to their propagation, meaning that systems located further north tend to have already spent some time over land. High values of mean_v500 and mean_v850 tend to push LPSs towards the Himalayas, where they rapidly dissipate, leading to smaller values of total_land_time. Higher values of mean_vort_850 also support greater inland penetration, as stronger systems can survive for longer once their energy source is removed. This is supported to a large extent by available barotropic instability (ushear_850_background) and to a lesser extent by midtropospheric thermodynamic instability (mean_dthetae_dp_750_500).

There is one relationship here that is, however, quite counter-intuitive. The total time LPSs spend over land is inversely proportional to the total column water vapour (TCWV) averaged over the monsoon core zone (mcz_tcwv); however, we would naively expect that a moister troposphere would provide a better environment for longer-lasting LPSs by supporting deeper and more widespread convection. We investigate this relationship further by plotting actual values of mcz_tcwv against their corresponding Shapley values in Figure 11a. Each point is coloured by the simultaneous value of mean_u200, which is a good proxy for vertical wind shear.

The response is non-monotonic and comprises three distinct components. For low values of mcz_tcwv($<50\,\mathrm{kg\cdot m^{-2}}$), there are two regimes: high shear (i.e., large negative values of mean_u200) and low shear. The high-shear regime is more than an order of magnitude less common, since strong vertical wind shear supports organised convection (Weisman and Klemp, 1982), even up to the smaller side of the synoptic scale of interest in our study (Baidu *et al.*, 2022). Points in the low-shear regime have a weak positive correlation with their respective Shapley values, implying that (when mcz_tcwv is low) increasing it generally supports longer LPS lifetime. For intermediate values of mcz_tcwv (between 50 and $65\,\mathrm{kg\cdot m^{-2}}$), which comprise the majority of cases, there is no longer a split between high- and low-shear cases. Instead, there is a transition, with increasing values of mcz_tcwv becoming increasingly associated with cases of strong vertical wind shear. In this intermediate regime, there is a negative correlation between mcz_tcwv and its Shapley values, meaning that increasing TCWV is associated with shorter LPS lifetime. This is the relationship brought out in Figure 10. Very high values of mcz_tcwv($>65\,\mathrm{kg\cdot m^{-2}}$) are dominated by large vertical wind shear. Here, the correlation between TCWV and its

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

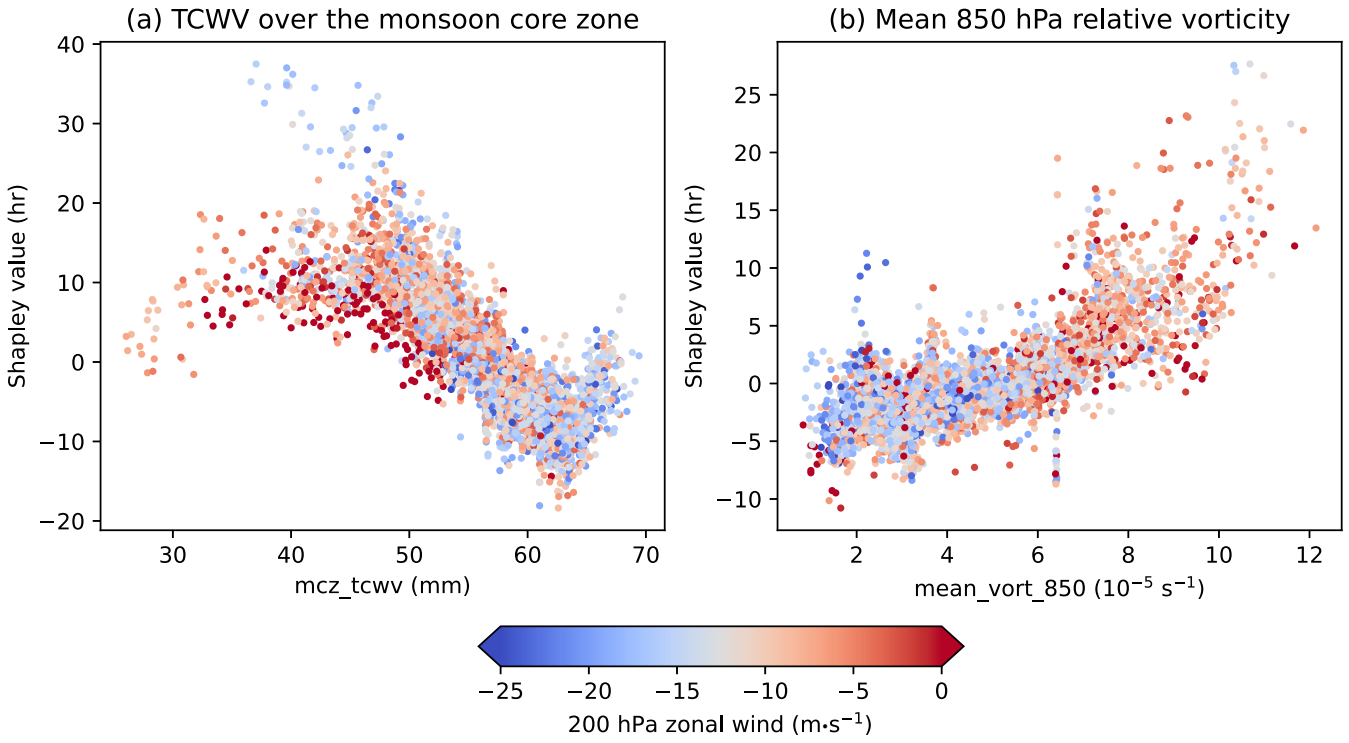

FIGURE 11 Relationship between (a) mcz_tcwv (total column water vapour [TCWV] over the monsoon core zone) and (b) mean_vort_850 (mean 850 hPa relative vorticity within 400 km of the low-pressure system centre) and their respective Shapley values in the total_land_time model. Each point in the bivariate distributions is coloured according to the simultaneous value of mean_u200 (mean 200 hPa zonal wind within 400 km of the low-pressure system centre).

Shapley values switches sign again—the effect of wind shear is essentially saturated, and so adding more moisture at this point helps the LPS survive for longer in a hostile environment.

Similar analysis with mean_vort_850 (Figure 11b) confirms the detrimental role that strong vertical wind shear plays in inland penetration of LPSs. Almost all cases of large vertical wind shear are associated with weaker LPSs, whereas LPSs occurring during weak shear conditions may be either weak or strong. We also note that predicted total_land_time is relatively insensitive to changes in mean_vort_850 for weak LPSs but grows quickly with increasing vorticity past about 6×10^{-5} s^{-1}.

Previous studies have suggested an important role for antecedent soil moisture in supporting post-landfall LPS durations (Baisya *et al.*, 2017; Hunt and Turner, 2017a). Following a similar analysis method as for the relationship between mcz_tcwv and mean_u200 (not shown), we find that for relatively dry columns (mcz_tcwv < 55 mm) that Shapley values of mcz_tcwv are positively correlated with soil moisture, indicating that soil moisture may indeed play an important role if the atmosphere above the trough is insufficiently humid. This relationship is not as strong as for vertical wind shear but indicates that land-surface feedbacks should not be neglected.

In summary, the surprising negative correlation between mcz_tcwv and total_land_time occurs because large vertical wind shear is required to support the organised convection that drives high values of TCWV over the monsoon core zone; but at the same time, for the majority of cases, such shear prevents strong LPSs from forming.

4.2.2 | Peak intensity

During our analysis of the previous model, we touched on what predictors affect LPS intensity. We expand on that in this section, where we predict the maximum intensity an LPS will reach during its lifetime, defined as the maximum value of mean_vort_850 it achieves. The model performs very well (Figure 12a), explaining two-thirds of the variance; but, as with the other models, it tends to underestimate high values and overestimate low values of the predictand.

Both lower tropospheric (mean_vort_850) and midtropospheric (mean_vort_500) vorticity are key predictors in the model. This is reasonably intuitive in a Bayesian framework: LPSs that attain a greater peak_vorticity have lifetime vorticity distributions that contains a greater fraction of high vorticity values than LPSs that only achieve a low

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

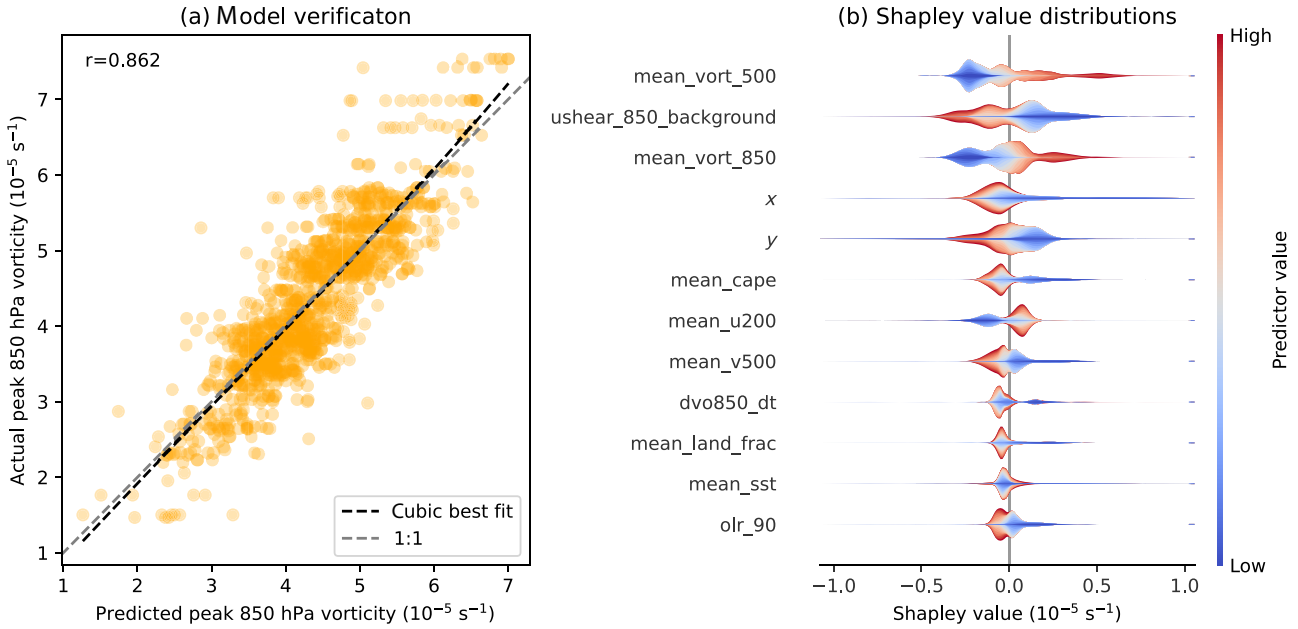

FIGURE 12 Verification and interpretation of the decision-tree model predicting peak_vorticity (the maximum value of mean_vort_850 reached during the LPS lifetime). (a) model predictions are plotted against observed values with the grey dashed line denoting a 1:1 relationship and the black dashed line showing a cubic best fit. The linear correlation coefficient between the actual and predicted values is given in the top left. (b) interpretation of the relative importance of predictors in the model is shown through their Shapley value distributions. The predictor variables are sorted by the mean of their absolute Shapley values with the distributions coloured according to the underlying value of the variable. This model converged after 111 rounds.

peak_vorticity. Therefore, a given LPS with high vorticity is more likely to attain a higher peak_vorticity than an LPS with low vorticity. It is interesting that, even though we measure peak_vorticity at 850 hPa, the model considers mean_vort_500 a more important predictor than mean_vort_850. This is probably because mean_vort_850 is a noisier field, being more vulnerable to feedbacks from the land surface and boundary layer and, as we saw in Section 4.1.1, the diurnal cycle.

The inclusion of x, y, and indeed mean_land_frac are also quite intuitive. For example, LPSs weaken considerably as they move further inland (decreasing x); so, for two LPSs of given vorticity, the one situated further west is likely to have had the greater peak_vorticity. We note, however, that the Shapley values are not particularly sensitive to changes in x for LPSs situated in the east (i.e., over the Bay of Bengal). This implies that LPSs do not necessarily intensify further by spending more time over the ocean.

The presence of ushear_850_background as the predictor with the second highest impact in the model supports the Diaz and Boos (2019a) theory of barotropic growth (noting that, like mean_u200, we expect the effect to be large for highly negative values, which here indicate a strong monsoon trough). So, whereas we saw in Section 4.1.1 that ushear_850_background was not a useful predictor of instantaneous LPS intensification (which responded more strongly to vertical wind shear, the land surface, and the diurnal cycle of convection), it is clearly useful in determining the maximum intensity LPSs are likely to reach.

We can investigate the relative impact of mean_vort_500 and ushear_850_background on model predictions as a function of x and y by plotting the mean Shapley value magnitudes on maps (Figure 13). We see that the model sensitivity to these parameters changes significantly as a function of LPS location, with mean_vort_500 typically having a higher impact over land and ushear_850_background over the ocean. In fact, the line that describes the mean longitude at which LPSs reach their maximum intensity partitions the two regions well. We infer from this that ushear_850_background is a useful predictor of potential LPS intensity, whereas mean_vort_500 is more useful in determining whether peak intensity has occurred and what its value was for decaying LPSs. We can confirm this by retraining the model only using points from LPSs that have not yet reached maximum intensity (not shown). That model performs equally well ($r = 0.83$) and has a similar order for predictor importance, except that ushear_850_background and mean_vort_500 swap places, with the former having a much greater impact than in the full model, and mean_land_frac moves up into third place. Combined, these results support our initial analysis in the second and third paragraphs of this subsection.

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

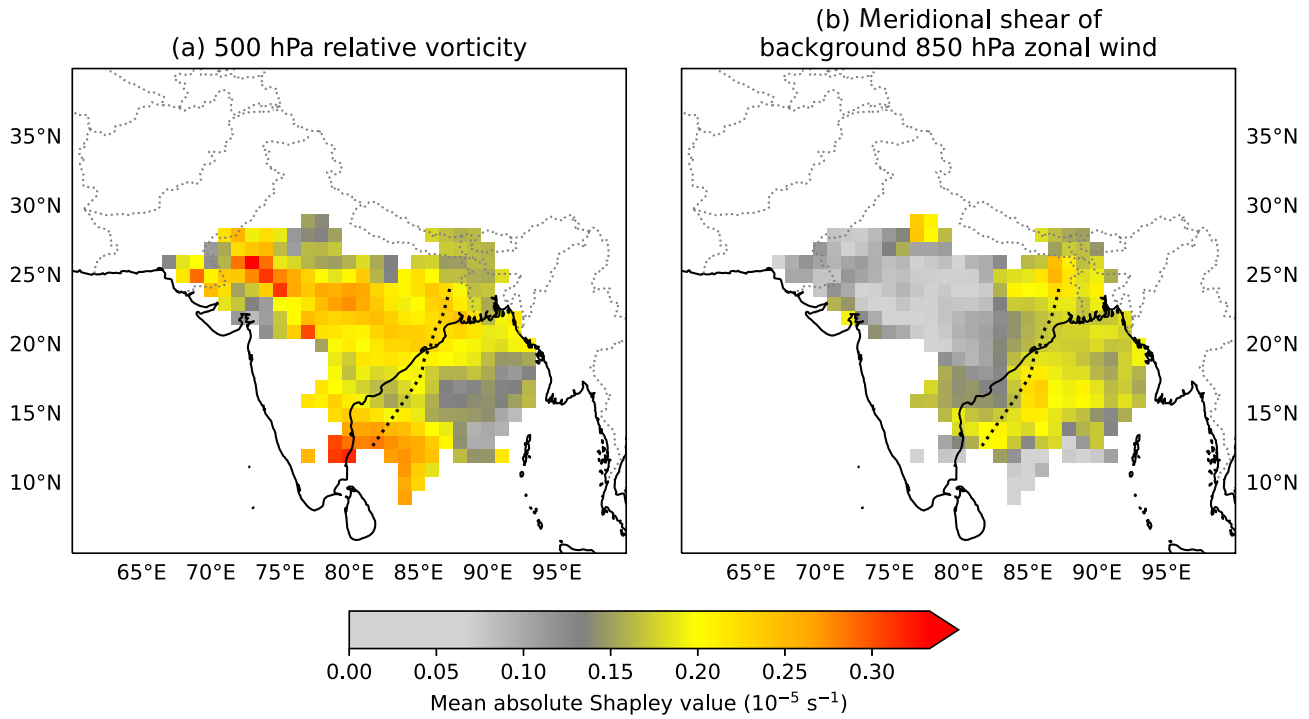

FIGURE 13 Map of mean Shapley value magnitudes for (a) mean_vort_500 (the mean value of 500 hPa vorticity within 400 km of the low-pressure system [LPS] centre) and (b) ushear_850_background (the mean value of large-scale meridional shear in 850 hPa zonal wind over the previous 10 days) in the peak_vorticity model. Data are binned onto a 1° × 1° grid according to the location of the LPS centre. Grid squares with fewer than five LPSs are not displayed. The dotted black line marks the approximate longitude where LPSs reach their peak vorticity, as a function of latitude, computed using lowess regression.

There are a few more predictors with reasonably high impact in Figure 12 that are worth discussing briefly. First, we note that mean_u200 is present with the same sign as in the total_land_time model, confirming the results there that large vertical wind shear ultimately suppresses the formation of strong LPSs. Second, sea-surface temperature (SST) only plays a role in some edge cases. This is surprising given that high SSTs have long been known to be an important precursor for LPS genesis (Sikka, 1977) and so presumably play some role in their early intensification. This is further confounded by the model relationship having the wrong sign (with stronger LPSs generally being associated with cooler SSTs). This may be explained by considering the behaviour of the monsoon trough. As we have seen, a strong trough supports sustained LPS intensification; however, such a trough is also associated with cooler SSTs over the Bay of Bengal due to increased surface runoff (Spiro Jaeger and Mahadevan, 2018) and due to increased upwelling forced by enhanced westerlies (Shetye *et al.*, 1991). This is clearly an area where more research is needed. Finally, we note another counter-intuitive relationship: predicted peak_vorticity tends to decrease with increasing mean_cape. It may be that LPSs with very high CAPE are already near their peak, as this instability is driven by strong quasi-geostrophic forcing and increased moisture content—this is supported by a similar but weaker relationship with olr_90—and therefore LPSs associated with low CAPE have more time to grow. Testing this hypothesis is left for future work.

4.2.3 | Precipitation

Our final model is trained to predict the mean precipitation falling within 400 km of the LPS centre over the next 6 hr. This model does not directly account for movement of the LPS, meaning that the area over which the average is computed is fixed for each training/testing point. This model performs very well (Figure 14a), accounting for almost three-quarters of the variance in near-term precipitation. This means that we can be particularly confident in having chosen a useful set of predictors and, therefore, in the results of the Shapley value analysis (Figure 14b).

The two mean vorticity terms (mean_vort_850 and mean_vort_500) have a high impact in the model, as expected. High values of either (or both) denote a more intense LPS and hence stronger quasi-geostrophic ascent, the primary forcing for large-scale precipitation associated with LPSs (Rajamani and Rao, 1981). Large mean_vort_850 also implies stronger surface winds,

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

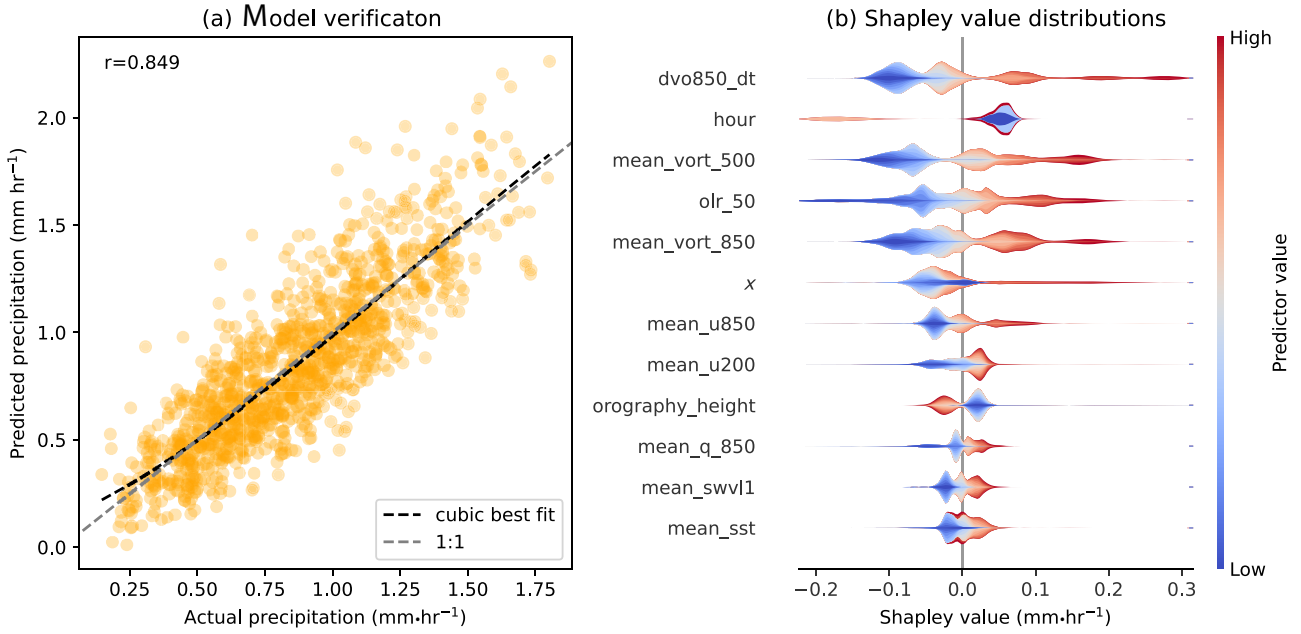

FIGURE 14 Verification and interpretation of the decision-tree model predicting mean_prcp_400 (the mean precipitation rate within 400 km of the low-pressure system centre over the following 6 hr). (a) Model predictions are plotted against observed values, with the grey dashed line denoting a 1:1 relationship and the black dashed line showing a cubic best fit. The linear correlation coefficient between the actual and predicted values is given in the top left. (b) Interpretation of the relative importance of predictors in the model is shown through their Shapley value distributions. The predictor variables are sorted by the mean of their absolute Shapley values, with the distributions coloured according to the underlying value of the variable. This model converged after 112 rounds.

increasing evaporation, and, thus, precipitation. This is corroborated to some extent by the relatively high impacts of mean_u850 and mean_sst. mean_vort_500 may also be more directly linked to deep convective rainfall through latent-heating-driven mid-level convergence. However, since this would be precipitation-driven vortex stretching (and hence larger vorticity), the causality would be in the wrong direction for rainfall prediction. High values of mean_vort_500 would indicate that the LPS is currently producing heavy rainfall. Similarly, the high impact of olr_50 shows that near-term rainfall is largely produced by existing deep convection, established over the large scale.

However, dvo850_dt has the highest impact of any predictor, indicating that an intensifying LPS is likely to rain more than a weakening one. This is not because stronger LPSs rain more (this relationship is picked out by mean_vort_850), but because the two are directly linked through vortex stretching. This would be a form of the moisture-vortex instability described by Adames and Ming (2018) and suggests that parametrising this process as we do in qshear_850_background (following Suhas and Boos, 2023) may not be appropriate.

The second highest impact comes from hour, again highlighting the importance of the diurnal cycle in the internal moist thermodynamics of LPSs, as we saw in Section 4.1.1. The large-scale afternoon minimum was also reported in Hunt *et al.* (2016b) and Hunt and Turner (2017b). Likewise, as in both the peak_vorticity and dvo850_dt models, strong vertical shear (seen here as highly negative mean_u200) inhibits LPS activity.

The absence of CAPE or stability terms in this model is notable. If we were measuring CAPE and precipitation simultaneously we might expect no relationship, or even a weakly negative correlation, since the convective precipitation associated with LPSs rapidly depletes CAPE. However, since we use the mean precipitation over the following 6 hr we would naively expect a positive correlation with CAPE, which leads precipitation by 2–4 hr in the Indian summer monsoon (Subrahmanyam *et al.*, 2015). In fact, of any such stability terms, predicted precipitation is only really dependent on readily available boundary-layer moisture (mean_q_850), and even then only weakly so. We therefore conclude that rainfall depends more on LPS dynamics than the suitability of the environment for convection.

Let us investigate those dynamics now, by exploring the relationship between LPS vortex intensity, LPS growth/decay, and precipitation in Figure 15. Here, each LPS time point is positioned according to its mean_vort500 and dvo850_dt and coloured according to its mean_prcp_400.

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

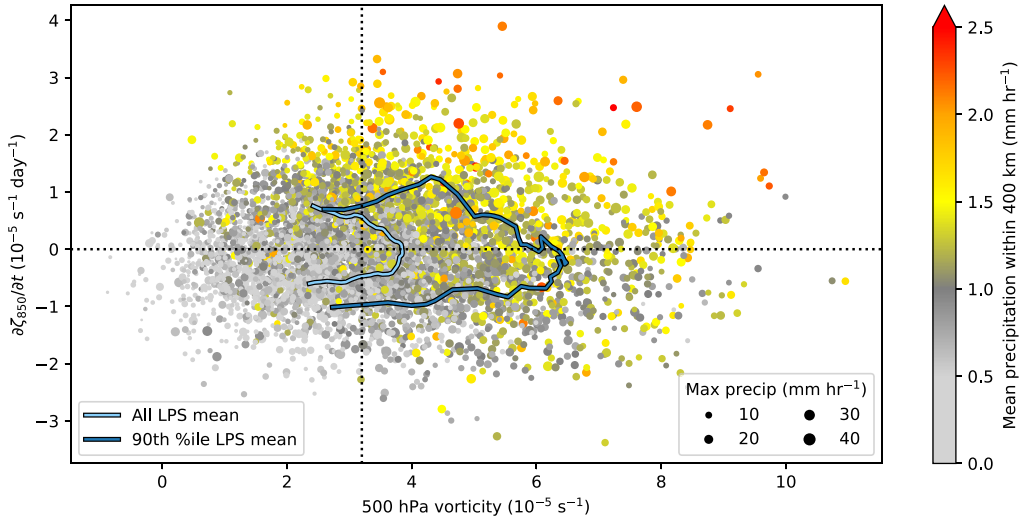

FIGURE 15 Scatter plot of dvo850_dt (the rate of change of mean 850 hPa relative vorticity within 400 km of the ow-pressure system [LPS] centre) against mean_vort_500 (the mean value of 500 hPa vorticity within 400 km of the LPS centre), for all LPS time points used in this study. Points are coloured according to mean_prcp_400 (the mean precipitation rate within 400 km of the LPS centre over the following 6 hr), and their radius is proportional to max_prcp_400 (as mean_prcp_400, using the mean precipitation over the next 6 hr, but instead taking the spatial maximum). Thick lines denote the mean trajectory in this phase space of (light blue) all LPSs and (darker blue) LPSs whose peak vorticity exceeds the 90th percentile. The vertical dotted line denotes the median value of mean_vort_500.

In addition, the size of the points scales with the maximum six-hourly precipitation rate within 400 km of the LPS centre. The mean LPS trajectory in this phase space starts in the top-left quadrant, moves clockwise, and finishes in the bottom-left quadrant. We see that weak and decaying LPSs (bottom-left quadrant, mostly over land) are very rarely associated with heavy rainfall. In contrast, if an LPS is both strong and intensifying (upper right quadrant) it is very likely to be associated with widespread heavy rainfall. However, individual extreme rainfall events are much less dependent on whether the LPS is growing or decaying and are in fact approximately equally likely to occur in any intense LPS.

5 | CONCLUSIONS

LPSs are the primary mode of synoptic-scale variability in the South Asian monsoon and bring the majority of its seasonal (Hunt and Fletcher, 2019) and extreme (Thomas *et al.*, 2021) precipitation. Although much research has been done on LPSs, especially on their stronger variety, known as monsoon depressions, important questions about their development and interaction with the environment and land and sea surface remained hitherto essentially unsolved. For example: What causes immature LPSs to intensify? Is it moist barotropic instability (Diaz and Boos, 2019b), convection-driven vortex stretching (Adames and Ming, 2018), or even CISK (Shukla, 1978)? What causes the average LPS to propagate northwestward when the low-level monsoon circulation is largely westerly over their domain? Is it beta drift of their mid-level PV maxima (Boos *et al.*, 2015), or off-centre vortex stretching (Goswami, 1987; Chen *et al.*, 2005), or even through interaction with the Himalayas (Hunt and Parker, 2016)? Is strong vertical wind shear detrimental to or supportive of LPS growth? There is weak evidence that LPSs are less likely to form in regions of vertical wind shear (Ditchek *et al.*, 2016), and such shear is well known to suppress tropical cyclone activity (DeMaria and Kaplan, 1994), but is vital for the development of organised convection within the monsoon (Weisman and Klemp, 1982). Given their hydrological, and hence societal, importance, how can we improve forecasts of LPS activity of, for example, their precipitation, peak intensity, and post-landfall behaviour?

In this article, we interrogated these dilemmata by posing them as parts of large statistical models known as additive decision-tree ensembles. We trained six such models, one each for LPS intensification rate, LPS zonal propagation speed, LPS meridional propagation speed, total time spent over land, peak LPS intensity, and near-term precipitation. Each model was trained using the XGBoost algorithm over about 40 predictor variables, which were drawn from earlier studies and pruned down using both correlation and clustering analysis to avoid cross-contamination during subsequent impact analysis. For each trained model, we then used Shapley value analysis to compute the relative impact of each predictor on the model predictions. We partition

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the subsequent discussion into new results and new research ideas.

5.1 | Summary of novel results

- Vertical wind shear exerts a weak control on LPS intensification. However, it has a considerably larger impact on peak LPS intensity and time spent over land, with our results indicating that strong vertical wind shear strongly inhibits further growth beyond a given LPS intensity.
- LPS intensification rate has a pronounced diurnal cycle, which varies depending on whether the system is over land or ocean. This invariably includes a strong early morning peak in growth, similar to that seen for precipitation in tropical cyclones and monsoon depressions. If the mechanism responsible is the same as for the precipitation peak, then it is because overnight radiative cooling of the upper level clouds causes atmospheric instability. Further research is needed to confirm this. LPS intensification is also sensitive to interaction with the land surface.
- The processes supporting (or suppressing) intensification differ from those supporting (or suppressing) overall peak intensity. The former is supported by processes acting over short time-scales (e.g., convection), whereas the latter is supported by processes acting over long time-scales (e.g., barotropic growth). This contrast is something that future modelling studies should take into account.
- Following Boos *et al.* (2015), we show that 500 hPa winds are a good predictor of LPS propagation speed and direction. However, this only holds so long as the LPSs are reasonably deep; that is, they possess a mid-level PV maximum that is comparable to or stronger than the low-level maximum. If this is not the case, and the LPS vorticity is largely confined to the lower troposphere, then LPS propagation is more strongly governed by low-level winds. We also found that both OLR and mid-level thermodynamic instability terms were important, suggesting a secondary role for vortex stretching, as proposed by Hunt and Turner (2022).
- More intense LPSs survive for longer over land, with post-landfall longevity increased by both mid-level thermodynamic instability and large-scale barotropic instability. However, TCWV over the monsoon core zone is, counter-intuitively, detrimental to LPS survival. This is because the vertical wind shear typically needed to maintain widespread organised convection within the monsoon tends to prevent strong LPSs from forming.
- Large-scale barotropic instability is a very good predictor of peak intensity (quantified using 850 hPa relative vorticity at the system centre) for LPSs currently over the ocean; however, mid-level vorticity (averaged within 400 km of the centre) is more useful for LPSs over land.
- Widespread short-term precipitation requires an LPS that is both intense and intensifying. However, individual extreme rainfall events only need an intense LPS and are not sensitive to its growth rate. Heavier rainfall is more likely in LPSs with well-established widespread convection.

5.2 | Summary of new research questions

Aside from testing existing hypotheses about LPS behaviour, our methodology also allows us to quickly investigate the relative importance of a large number of variables for LPS processes. From this, we can generate hypotheses and research questions. These are listed in the following.

- The diurnal cycle of LPS intensification is aligned with the diurnal cycle of convection over the ocean but not over land. Does this play a role in hastening LPS demise post-landfall?
- LPS peak intensity is not sensitive to the time spent over the ocean, nor particularly to underlying SSTs. However, it has been known for a long time that SSTs play an important role in LPS genesis (Sikka, 1977). So, what is the relationship between early LPS growth and the sea surface? Are SST gradients important? Is there a negative feedback between a strong monsoon trough (good for growth through barotropic instability) and cooling of SSTs due to westerly induced coastal upwelling off the east coast of India (bad for growth through surface heat exchange mechanisms such as wind-induced surface heat exchange)?
- Similarly, low-level wind speed is a good predictor of short-term LPS precipitation; however, the associated convection also intensifies the LPS through vortex stretching. Is this evidence that LPSs can intensify through wind-induced surface heat exchange?
- Why is there an inverse relationship between CAPE (and OLR) and predicted peak LPS intensity? Is it simply because LPSs associated with high CAPE or low OLR are already near their peak, or is a more complex interaction at play?
- Increased deep convection leads to more mid-level latent heating. This, in turn, strengthens the negative

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

anomaly in midtropospheric geopotential. How important is this feedback for LPSs?

AUTHOR CONTRIBUTIONS

Kieran M. R. Hunt: conceptualization; formal analysis; investigation; methodology; software; validation; visualization; writing – original draft; writing – review and editing. **Andrew G. Turner:** conceptualization; funding acquisition; investigation; project administration; writing – review and editing.

ACKNOWLEDGEMENTS

Kieran M. R. Hunt and Andrew G. Turner were funded for this work through the Weather and Climate Science for Service Partnership (WCSSP) India, a collaborative initiative between the Met Office, supported by the UK Government's Newton Fund, and the Indian Ministry of Earth Sciences (MoES).

DATA AVAILABILITY STATEMENT

All the code needed to reproduce the models, analysis, and figures for this article is available at https://github.com/kieranmrhunt/lps-xgboost. The full set of ERA5 LPS tracks is available at https://doi.org/10.5281/zenodo.7568990. The filtered ERA5 LPS track data, including the additional environmental data described in Section 3.1, are available at https://doi.org/10.5281/zenodo.7569057.

ORCID

Kieran M. R. Hunt https://orcid.org/0000-0003-1480-3755
Andrew G. Turner https://orcid.org/0000-0002-0642-6876

REFERENCES

Adames, Á.F. (2021) Interactions between water vapor, potential vorticity, and vertical wind shear in quasi-geostrophic motions: implications for rotational tropical motion systems. *Journal of the Atmospheric Sciences*, 78, 903–923.

Adames, Á.F. and Ming, Y. (2018) Interactions between water vapor and potential vorticity in synoptic-scale monsoonal disturbances: moisture vortex instability. *Journal of the Atmospheric Sciences*, 75, 2083–2106.

Baidu, M., Schwendike, J., Marsham, J.H. and Bain, C. (2022) Effects of vertical wind shear on intensities of mesoscale convective systems over west and Central Africa. *Atmospheric Science Letters*, 23, e1094.

Baisya, H., Pattnaik, S. and Rajesh, P.V. (2017) Land surface-precipitation feedback analysis for a landfalling monsoon depression in the Indian region. Accepted.

Barton, E.J., Taylor, C.M., Parker, D.J., Turner, A.G., Belušić, D., Böing, S.J., Brooke, J.K., Harlow, R.C., Harris, P.P. and Hunt, K. (2020) A case-study of land–atmosphere coupling during monsoon onset in northern India. *Quarterly Journal of the Royal Meteorological Society*, 146, 2891–2905.

Boos, W.R., Hurley, J.V. and Murthy, V.S. (2015) Adiabatic westward drift of Indian monsoon depressions. *Quarterly Journal of the Royal Meteorological Society*, 141, 1035–1048. https://doi.org/10.1002/qj.2454.

Bosilovich, M.G., Chen, J., Robertson, F.R. and Adler, R.F. (2008) Evaluation of global precipitation in reanalyses. *Journal of Applied Meteorology and Climatology*, 47, 2279–2299. https://doi.org/10.1175/2008JAMC1921.1.

Bowman, K.P. and Fowler, M.D. (2015) The diurnal cycle of precipitation in tropical cyclones. *Journal of Climate*, 28, 5325–5334. https://doi.org/10.1175/JCLI-D-14-00804.1.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (2017) *Classification and Regression Trees*. New York: Routledge.

Brochu, E., Cora, V.M. and De Freitas, N. (2010) A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv preprint arXiv:1012.2599.

Chang, H.I., Niyogi, D., Kumar, A., Kishtawal, C.M., Dudhia, J., Chen, F., Mohanty, U.C. and Shepherd, M. (2009) Possible relation between land surface feedback and the post-landfall structure of monsoon depressions. *Geophysical Research Letters*, 36. https://doi.org/10.1029/2009GL037781.

Chen, T. and Guestrin, C. (2016) Xgboost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 785–794.

Chen, T.-C., Yoon, J.-H. and Wang, S.-Y. (2005) Westward propagation of the Indian monsoon depression. *Tellus*, 57A, 758–769. https://doi.org/10.1111/j.1600-0870.2005.00140.x.

Cohen, N.Y. and Boos, W.R. (2016) Perspectives on moist baroclinic instability: implications for the growth of monsoon depressions. *Journal of the Atmospheric Sciences*, 73, 1767–1788. https://doi.org/10.1175/JAS-D-15-0254.1.

C-SharpCorner. (2021) Xgboost-the choice of most champions. https://www.c-sharpcorner.com/article/xgboost-the-choice-of-most-champions.

Dataaspirant. (2020) How the kaggle winners algorithm xgboost algorithm works. https://dataaspirant.com/xgboost-algorithm.

DeMaria, M. and Kaplan, J. (1994) A statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic basin. *Weather and Forecasting*, 9, 209–220.

Deoras, A., Turner, A.G. and Hunt, K.M.R. (2022) The structure of strong Indian monsoon low-pressure systems in subseasonal-to-seasonal prediction models. *Quarterly Journal of the Royal Meteorological Society*, 148, 2147–2166.

Diaz, M. and Boos, W.R. (2019a) Barotropic growth of monsoon depressions. *Quarterly Journal of the Royal Meteorological Society*, 145, 824–844.

Diaz, M. and Boos, W.R. (2019b) Monsoon depression amplification by moist barotropic instability in a vertically sheared environment. *Quarterly Journal of the Royal Meteorological Society*, 145, 2666–2684.

Diaz, M. and Boos, W.R. (2021) Evolution of idealized vortices in monsoon-like shears: application to monsoon depressions. *Journal of the Atmospheric Sciences*, 78, 1207–1225.

Ditchek, S.D., Boos, W.R., Camargo, S.J. and Tippett, M.K. (2016) A genesis index for monsoon disturbances. *Journal of Climate*, 29(14), 5189-5203. https://doi.org/10.1175/JCLI-D-15-0704.1.

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Dong, W.-H., Ming, Y. and Ramaswamy, V. (2020) Projected changes in south Asian monsoon low-pressure systems. *Journal of Climate*, 33, 7275–7287.

Friedman, J.H. (2001) Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Gillett, N., Allen, M. and Tett, S. (2000) Modelled and observed variability in atmospheric vertical temperature structure. *Climate Dynamics*, 16, 49–61.

Godbole, R.V. (1977) The composite structure of the monsoon depression. *Tellus*, 29, 25–40. https://doi.org/10.1111/j.2153-3490.1977.tb00706.x.

Goswami, B.N. (1987) A mechanism for the west-north-west movement of monsoon depressions. *Nature*, 326, 376–378. https://doi.org/10.1038/326376a0.

Goswami, B.N., Keshavamurty, R.N. and Satyan, V. (1980) Role of barotropic, baroclinic and combined barotropic-baroclinic instability for the growth of monsoon depressions and mid-tropospheric cyclones. *Proceedings of the Indian Academy of Sciences-Earth and Planetary Sciences*, 89, 79–97.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R. and Schepers, D. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049.

Hunt, K.M.R. and Fletcher, J.K. (2019) The relationship between Indian monsoon rainfall and low-pressure systems. *Climate Dynamics*, 53, 1–13.

Hunt, K.M.R. and Parker, D.J. (2016) The movement of Indian monsoon depressions by interaction with image vortices near the Himalayan wall. *Quarterly Journal of the Royal Meteorological Society*, 142, 2224–2229. https://doi.org/10.1002/qj.2812.

Hunt, K.M.R. and Turner, A.G. (2017a) The effect of soil moisture perturbations on Indian monsoon depressions in a numerical weather prediction model. *Journal of Climate*, 30, 8811–8823.

Hunt, K.M.R. and Turner, A.G. (2017b) The representation of Indian monsoon depressions at different horizontal resolutions in the met office unified model. *Quarterly Journal of the Royal Meteorological Society*, 143, 1756–1771. https://doi.org/10.1002/qj.3030.

Hunt, K.M.R. and Turner, A.G. (2022) Non-linear intensification of monsoon low-pressure systems by the BSISO. *Weather and Climate Dynamics*, 3, 1341–1358.

Hunt, K.M.R., Turner, A.G., Inness, P.M., Parker, D.E. and Levine, R.C. (2016a) On the structure and dynamics of Indian monsoon depressions. *Monthly Weather Review*, 144, 3391–3416. https://doi.org/10.1175/MWR-D-15-0138.1.

Hunt, K.M.R., Turner, A.G. and Parker, D.E. (2016b) The spatiotemporal structure of precipitation in Indian monsoon depressions. *Quarterly Journal of the Royal Meteorological Society*, 142, 3195–3210. https://doi.org/10.1002/qj.2901.

Hurley, J.V. and Boos, W.R. (2015) A global climatology of monsoon low pressure systems. *Quarterly Journal of the Royal Meteorological Society*, 141, 1049–1064. https://doi.org/10.1002/qj.2447.

IMD. (2003) Cyclone manual.

Jones, D.R., Schonlau, M. and Welch, W.J. (1998) Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13, 455–492.

Kishtawal, C., Niyogi, D., Rajagopalan, B., Rajeevan, M., Jaiswal, N. and Mohanty, U. (2013) Enhancement of inland penetration of monsoon depressions in the bay of Bengal due to prestorm ground wetness. *Water Resources Research*, 49, 3589–3600. https://doi.org/10.1002/wrcr.20301.

Kotsiantis, S.B. (2013) Decision trees: a recent overview. *Artificial Intelligence Review*, 39, 261–283.

Krishnakumar, V., Keshavamurty, R.N. and Kasture, S.V. (1992) Moist baroclinic instability and the growth of monsoon depressions–linear and nonlinear studies. *Proceedings of the Indian National Science Academy*, 101, 123–152.

Liu, C. and Zipser, E.J. (2008) Diurnal cycles of precipitation, clouds, and lightning in the tropics from 9 years of TRMM observations. *Geophysical Research Letters*, 35.

Loh, W.-Y. (2011) Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1, 14–23.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.-I. (2020) From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, 56–67.

Lundberg, S.M. and Lee, S.-I. (2017) A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R. (Eds.) *Advances in Neural Information Processing Systems*, Vol. 30. Long Beach, CA: Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Ma, L., Zhang, T., Frauenfeld, O.W., Ye, B., Yang, D. and Qin, D. (2009) Evaluation of precipitation from the ERA-40, NCEP-1, and NCEP-2 reanalyses and CMAP-1, CMAP-2, and GPCP-2 with ground-based measurements in China. *Journal of Geophysical Research: Atmospheres*, 114, D09105.

Mamgain, A., Rajagopal, E.N., Mitra, A.K. and Webster, S. (2018) Short-range prediction of monsoon precipitation by NCMRWF regional unified model with explicit convection. *Pure and Applied Geophysics*, 175, 1197–1218.

Martin, G.M., Brooks, M.E., Johnson, B., Milton, S.F., Webster, S., Jayakumar, A., Mitra, A.K., Rajan, D. and Hunt, K.M.R. (2020) Forecasting the monsoon on daily to seasonal time-scales in support of a field campaign. *Quarterly Journal of the Royal Meteorological Society*, 146, 2906–2927.

Mockus, J. (1994) Application of Bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, 4, 347–365.

Moorthi, S. and Arakawa, A. (1985) Baroclinic instability with cumulus heating. *Journal of the Atmospheric Sciences*, 42, 2007–2031. https://doi.org/10.1175/1520-0469(1985)042<2007:BIWCH>2.0.CO;2.

Nitta, T. and Masuda, K. (1981) Observational study of a monsoon depression developed over the bay of Bengal during summer MONEX. *Journal of the Meteorological Society of Japan*, 59, 672–682.

Podeti, S.R., Ramakrishna, S.S.V.S., Viswanadhapalli, Y., Dasari, H., Nellipudi, N.R. and Rao, B. (2020) Sensitivity of cloud microphysics on the simulation of a monsoon depression over the bay of Bengal. *Pure and Applied Geophysics*, 177, 5487–5505.

Quinlan, J.R. (1986) Induction of decision trees. *Machine Learning*, 1, 81–106.

Rajamani, S. and Rao, K.V. (1981) On the occurrence of rainfall over southwest sector of monsoon depression. *Mausam*, 32, 215–220.

Rajamani, S. and Sikdar, D.N. (1989) Some dynamical characteristics and thermal structure of monsoon depressions over the bay of Bengal. *Tellus A*, 41, 255–269.

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Roth, A.E. (1988) *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge: Cambridge University Press.

Roy, P. and Rao, T.N. (2022) Precipitation characteristics of cyclonic disturbances over South Asia region as revealed by TRMM and GPM. *Journal of Climate*, *35*(15), 4943-4957.

Salvekar, P.S., George, L. and Mishra, S.K. (1986) Low level wind shear and baroclinic growth of monsoon depression scale waves. *Meteorology and Atmospheric Physics*, 35, 10–18.

Sanders, F. (1984) Quasi-geostrophic diagnosis of the monsoon depression of 5-8 July 1979. *Journal of the Atmospheric Sciences*, 41, 538–552. https://doi.org/10.1175/1520-0469(1984)041<0538:QGDOTM>2.0.CO;2.

Shapley, L. (1953) A value for n-person games. In Kuhn, H. and Tucker, A., Eds., *Contributions to the Theory of Games II*. Princeton: Princeton University Press, 307-317.

Shetye, S.R., Shenoi, S.S.C., Gouveia, A.D., Michael, G.S., Sundar, D. and Nampoothiri, G. (1991) Wind-driven coastal upwelling along the western boundary of the bay of Bengal during the southwest monsoon. *Continental Shelf Research*, 11, 1397–1408.

Shukla, J. (1978) CISK-barotropic-baroclinic instability and the growth of monsoon depressions. *Journal of the Atmospheric Sciences*, 35, 495–508.

Sikka, D.R. (1977) Some aspects of the life history, structure and movement of monsoon depressions. *Pure and Applied Geophysics*, 115, 1501–1529. https://doi.org/10.1007/BF00874421.

Sikka, D.R. (2006) *A Study on the Monsoon Low Pressure Systems over the Indian Region and their Relationship with Drought and Excess Monsoon Seasonal Rainfall*. Fairfax, VA: Center for Ocean-Land-Atmosphere Studies, Center for the Application of Research on the Environment.

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S. and Morcos, A.S. (2022) Beyond neural scaling laws: beating power law scaling via data pruning. arXiv preprint arXiv:2206.14486.

Spiro Jaeger, G. and Mahadevan, A. (2018) Submesoscale-selective compensation of fronts in a salinity-stratified ocean. *Science Advances*, 4, e1701504.

Stano, G., Krishnamurti, T.N., Vijaya Kumar, T.S.V. and Chakraborty, A. (2002) Hydrometeor structure of a composite monsoon depression using the TRMM radar. *Tellus A*, 54, 370–381. https://doi.org/10.1034/j.1600-0870.2002.01330.x.

Subrahmanyam, D., Tandon, M.K., George, L. and Mishra, S.K. (1981) Role of barotropic mechanism in the development of a monsoon depression: a MONEX study. *Pure and Applied Geophysics*, 119, 901–912.

Subrahmanyam, K.V., Kumar, K.K. and Narendra Babu, A. (2015) Phase relation between CAPE and precipitation at diurnal scales over the Indian summer monsoon region. *Atmospheric Science Letters*, 16, 346–354.

Suhas, D.L. and Boos, W.R. (2023) Monsoon depression amplification by horizontal shear and humidity gradients: a shallow water perspective. *Journal of the Atmospheric Sciences,* 80(2), 633-647.

Taszarek, M., Pilguj, N., Allen, J.T., Gensini, V., Brooks, H.E. and Szuster, P. (2021) Comparison of convective parameters derived from ERA5 and MERRA-2 with rawinsonde data over Europe and North America. *Journal of Climate*, 34, 3211–3237.

Thomas, T.M., Bala, G. and Srinivas, V.V. (2021) Characteristics of the monsoon low pressure systems in the Indian subcontinent and the associated extreme precipitation events. *Climate Dynamics*, 56, 1859–1878.

Turner, A.G., Bhat, G.S., Martin, G.M., Parker, D.J., Taylor, C.M., Mitra, A.K., Tripathi, S.N., Milton, S., Rajagopal, E.N. and Evans, J.G. (2020) Interaction of convective organization with monsoon precipitation, atmosphere, surface and sea: the 2016 INCOMPASS field campaign in India. *Quarterly Journal of the Royal Meteorological Society*, 146, 2828–2852.

Vidhya, A. (2016) Winning solutions of dyd competition–r and xgboost ruled. https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python.

Vishnu, S., Boos, W.R., Ullrich, P.A. and O'Brien, T.A. (2020) Assessing historical variability of south Asian monsoon lows and depressions with an optimized tracking algorithm. *Journal of Geophysical Research: Atmospheres*, 125, e2020JD032977.

Wang, J., Clark, S.C., Liu, E. and Frazier, P.I. (2020) Parallel Bayesian global optimization of expensive functions. *Operations Research*, 68, 1850–1865.

Wang, Y. and Chen, Y. (2019) Significant climate impact of highly hygroscopic atmospheric aerosols in Delhi, India. *Geophysical Research Letters*, 46, 5535–5545.

Wang, Z., Franke, J.A., Luo, Z. and Moyer, E.J. (2021) Reanalyses and a high-resolution model fail to capture the "high tail" of CAPE distributions. *Journal of Climate*, 34, 8699–8715.

Weisman, M.L. and Klemp, J.B. (1982) The dependence of numerically simulated convective storms on vertical wind shear and buoyancy. *Monthly Weather Review*, 110, 504–520.

Xu, J., Ma, Z., Yan, S. and Peng, J. (2022) Do ERA5 and ERA5-land precipitation estimates outperform satellite-based precipitation products? A comprehensive comparison between state-of-the-art model-based and satellite-based precipitation products over mainland China. *Journal of Hydrology*, 605, 127353.

Yang, G.-Y. and Slingo, J. (2001) The diurnal cycle in the tropics. *Monthly Weather Review*, 129, 784–801.

How to cite this article: Hunt, K.M.R. & Turner, A.G. (2023) Using interpretable gradient-boosted decision-tree ensembles to uncover novel dynamical relationships governing monsoon low-pressure systems. *Quarterly Journal of the Royal Meteorological Society*, 1–24. Available from: https://doi.org/10.1002/qj.4582

1477870x, 0, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4582 by Test, Wiley Online Library on [18/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License