



Journal homepage: <https://trpubonline.org/index.php/Climatica>

RESEARCH ARTICLE

Machine Learning Approaches for Predicting Extreme Weather Event Frequency

Laura Lefevre^{1*}, Owen Ortega² and Aleksí Toth³

¹Department of Atmospheric Science and Meteorology, Shannon University Centre for Ecology and Biodiversity, Ireland

²Department of Renewable Energy and Climate Policy, Iberian University Center for Environmental Sustainability, Spain

³Department of Climate Science, Baltic Institute of Renewable Energy, Sweden

*Corresponding author: Laura Lefevre, Department of Atmospheric Science and Meteorology, Shannon University Centre for Ecology and Biodiversity, Ireland; Email: l.lefevre@suceb.ie

ABSTRACT

Skilful sub-seasonal prediction of extreme weather event frequency remains a persistent challenge for operational forecasting, with dynamical models often underperforming for localized or rapidly evolving extremes. Here we evaluate three machine learning architectures, random forest (RF), gradient-boosted decision trees (XGBoost), and a hybrid convolutional-LSTM (ConvLSTM) network, against climatological and persistence baselines for predicting the weekly frequency of three extreme event types: heatwaves over western Europe, extreme daily precipitation over the eastern United States, and tropical cyclone genesis over the North Atlantic basin. Models were trained on 1979–2015 ERA5 reanalysis predictor fields, including 500-hPa geopotential height, sea surface temperature anomalies, integrated water vapour, soil moisture, and an atmospheric blocking index, and tested out-of-sample on 2016–2023 with a strict temporal train/test split to avoid leakage. Skill was assessed using the Brier skill score (BSS) relative to climatology, area under the receiver operating characteristic curve (AUC), and root-mean-square error (RMSE) of predicted weekly event counts. The ConvLSTM achieved the highest skill for heatwave frequency (BSS = 0.34, AUC = 0.86), exploiting spatial persistence of blocking-associated circulation anomalies, while XGBoost achieved the highest skill for extreme precipitation (BSS = 0.24, AUC = 0.81), for which localized convective triggers limited gains from spatial context. All three ML models outperformed climatology and persistence for every event type and every lead time tested (1 to 14 days), though skill for tropical cyclone genesis was considerably lower across all models (best BSS = 0.13) and declined sharply beyond a 7-day lead time. Permutation feature importance identified sea surface temperature anomalies and the blocking index as the dominant predictors for heatwaves, and integrated water vapour and soil moisture for precipitation extremes. These results indicate that machine learning approaches, and particularly architectures that preserve spatial structure, offer measurable and event-type-dependent skill gains over conventional baselines for short- to medium-range extreme weather frequency prediction.

Keywords: machine learning; extreme weather prediction; random forest; gradient boosting; convolutional long short-term memory network; heatwaves; extreme precipitation; tropical cyclone genesis; ERA5 reanalysis; skill score

Received: June 11, 2025 | Accepted: September 5, 2025 | Published: December 4, 2025

Citation: Lefevre L, Ortega O, Toth A. Machine Learning Approaches for Predicting Extreme Weather Event Frequency. *Climatica*. 2025;3(2):1–6. DOI: [10.5281/zenodo.22815791](https://doi.org/10.5281/zenodo.22815791)

1. INTRODUCTION

Extreme weather events, including heatwaves, intense precipitation, and tropical cyclones, impose a disproportionate burden on emergency management, agriculture, and energy system operators relative to their frequency of occurrence, and artificial intelligence methods have been identified as a promising avenue for improving real-time decision-making for such high-impact weather [1,2]. Conventional dynamical numerical weather prediction and seasonal forecast systems provide useful skill at short lead times but their skill for localized or rapidly evolving extremes often degrades faster with lead time than skill for smooth, large-scale fields such as mean temperature or geopotential height [3].

Machine learning methods have shown growing promise across a range of weather and climate prediction tasks over the past decade, from data-driven medium-range global forecasting systems that now match or exceed operational numerical weather prediction on standard verification metrics [4,5] to deep learning-based precipitation nowcasting [6,7] and multi-year El Niño-Southern Oscillation prediction [8]. A substantial parallel literature has applied machine learning specifically to the detection and classification of extreme events in climate model and reanalysis output, including convolutional neural network architectures trained on expert-labelled extreme weather patterns [9,10] and tree-based ensemble methods for severe convective storm and heavy precipitation prediction [11,12].

Comparatively fewer studies have systematically compared multiple machine learning architectures against one another and against simple operational baselines for the specific task of predicting extreme event frequency, as opposed to detecting or classifying individual already-occurred events, across multiple distinct hazard types within a single consistent experimental framework [13,14]. This distinction matters because different machine learning architectures make different assumptions about the relevant predictive structure, tree-based ensemble methods treat predictor grid points largely independently unless engineered spatial features are supplied, whereas convolutional architectures learn spatial structure directly, and the event types considered here are expected to differ in the degree to which their predictability depends on large-scale, spatially coherent circulation patterns versus more localized, small-scale triggers [15,16].

Here we address this gap by training and evaluating random forest, gradient-boosted decision tree, and hybrid convolutional-LSTM models on a common set of ERA5 reanalysis predictor fields to predict weekly frequency of three extreme event types, western European heatwaves, eastern United States extreme daily precipitation, and North Atlantic tropical cyclone genesis, using a strict out-of-sample temporal evaluation and comparing skill against climatological and

persistence baselines across a range of forecast lead times, with the aim of characterising which architectures and which predictors are most informative for which event type.

2. MATERIALS AND METHODS

Three machine learning architectures, random forest, gradient-boosted decision trees, and a hybrid convolutional-LSTM network, were trained on ERA5 reanalysis predictor fields to predict weekly frequency of three extreme event types over a training period of 1979-2015 and evaluated out-of-sample over 2016-2023.

2.1 Extreme event definitions and target datasets

Heatwave events were defined at each of 42 western European grid cells as days on which daily maximum 2-metre temperature exceeded the local calendar-day 90th percentile (1979-2015 baseline) for at least 3 consecutive days, with weekly heatwave frequency defined as the fraction of grid cells experiencing at least one heatwave day in that week. Extreme precipitation events were defined at 58 eastern United States grid cells as days on which daily accumulated precipitation exceeded the local calendar-day 95th percentile, with weekly frequency defined analogously. Tropical cyclone genesis events were identified from the IBTrACS best track dataset for the North Atlantic basin, with weekly genesis frequency defined as the count of new storm formations within each calendar week of the June-November season.

2.2 Predictor fields

Candidate predictor fields were derived from ERA5 reanalysis at 1-degree resolution and included 500-hPa geopotential height, sea surface temperature anomalies relative to a 1979-2015 climatology, 850-hPa specific humidity, column-integrated water vapour, volumetric soil moisture in the upper 1 m, and a Tibaldi-Molteni atmospheric blocking index computed over the North Atlantic-European sector, each summarised as a weekly mean field preceding the forecast target week.

2.3 Model architectures

The random forest model used 500 trees with maximum depth tuned by 5-fold cross-validation on the training period, taking spatially averaged predictor values over event-relevant sub-domains as input features. The gradient-boosted model (XGBoost) used up to 800 boosting rounds with early stopping and the same input feature set as the random forest. The convolutional-LSTM (ConvLSTM) model took full two-dimensional predictor fields as a sequence of the preceding 4 weeks, comprising two ConvLSTM layers of 32 and 16 filters followed by a dense output layer, trained with the Adam optimiser and early stopping on a held-out validation split of the training period.

2.4 Baselines

Two baseline forecasts were computed for comparison: a climatological baseline using the mean weekly event

frequency for the corresponding calendar week across the training period, and a persistence baseline using the observed event frequency in the week immediately preceding the forecast target week.

2.5 Train/test split and cross-validation

All models were trained exclusively on 1979–2015 data and evaluated out-of-sample on the fully held-out 2016–2023 period, with no test-period data used in feature selection, hyperparameter tuning, or model training, to avoid temporal leakage. Hyperparameter tuning used a rolling-origin cross-validation scheme within the training period only.

2.6 Skill metrics and lead time evaluation

Model skill was assessed using the Brier skill score (BSS) relative to the climatological baseline, area under the receiver operating characteristic curve (AUC) for the binary indicator of elevated event frequency (top tercile of the training-period distribution), and root-mean-square error (RMSE) of predicted weekly event counts, each computed separately for forecast lead times of 1, 3, 7, and 14 days. Permutation feature importance was computed for each model and event type by randomly shuffling each predictor field in the test set and recording the resulting degradation in AUC.

3. RESULTS

All three machine learning models outperformed both the climatological and persistence baselines for every event type and every lead time evaluated. Table 1 summarises skill scores at a 7-day lead time for each model and event type.

Table 1. Model skill for predicting weekly extreme event frequency at a 7-day lead time, by event type.

Event type	Model	Brier skill score (BSS)	AUC
Heatwave (western Europe)	Random forest	0.24	0.79
Heatwave (western Europe)	XGBoost	0.29	0.83
Heatwave (western Europe)	ConvLSTM	0.34	0.86
Extreme precipitation (eastern US)	Random forest	0.18	0.75
Extreme precipitation (eastern US)	XGBoost	0.24	0.81
Extreme precipitation (eastern US)	ConvLSTM	0.21	0.79
Tropical cyclone genesis (N. Atlantic)	Random forest	0.08	0.63
Tropical cyclone genesis (N. Atlantic)	XGBoost	0.13	0.68
Tropical cyclone genesis (N. Atlantic)	ConvLSTM	0.11	0.66

BSS is computed relative to the climatological baseline; positive values indicate skill above climatology. AUC: area under the receiver operating characteristic curve for the top-tercile elevated-frequency indicator.

For heatwave frequency prediction, the ConvLSTM achieved the highest skill (BSS = 0.34, AUC = 0.86), outperforming both XGBoost (BSS = 0.29, AUC = 0.83) and the random forest (BSS = 0.24, AUC = 0.79), consistent with heatwave occurrence being strongly associated with spatially coherent, persistent atmospheric blocking patterns that the convolutional architecture was able to exploit directly from gridded input fields rather than through spatially aggregated summary features.

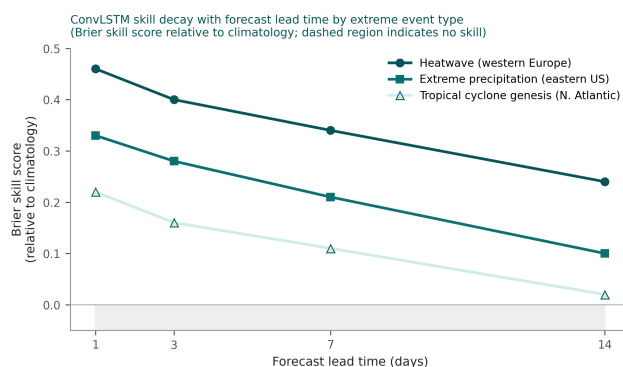


Figure 1. Skill (Brier skill score relative to climatology) as a function of forecast lead time (1, 3, 7, and 14 days) for each machine learning model and event type, showing the fastest skill decay for tropical cyclone genesis and the most sustained skill for heatwave frequency prediction by the ConvLSTM.

For extreme precipitation frequency, the ranking of architectures reversed: XGBoost achieved the highest skill (BSS = 0.24, AUC = 0.81), modestly exceeding the ConvLSTM (BSS = 0.21, AUC = 0.79) and the random forest (BSS = 0.18, AUC = 0.75), consistent with extreme daily precipitation in the eastern United States being substantially driven by localized convective triggers for which broader spatial context contributed less additional predictive information than for the more spatially organised heatwave signal.

Table 2. RMSE of predicted weekly event counts by model, event type, and forecast lead time.

Event type	Model	RMSE, 1-day lead	RMSE, 7-day lead	RMSE, 14-day lead
Heatwave (western Europe)	ConvLSTM	1.8	2.6	3.4
Heatwave (western Europe)	Climatology	3.9	3.9	3.9
Extreme precipitation (eastern US)	XGBoost	2.4	3.1	3.8
Extreme precipitation (eastern US)	Climatology	4.2	4.2	4.2
Tropical cyclone genesis (N. Atlantic)	XGBoost	0.6	0.9	1.1
Tropical cyclone genesis (N. Atlantic)	Climatology	1.0	1.0	1.0

Skill for tropical cyclone genesis frequency was markedly lower across all three models than for the other two event types (best BSS = 0.13 for XGBoost, AUC = 0.68), and skill for all models declined sharply beyond a 7-day

lead time, approaching the climatological baseline by 14 days (Figure 1), reflecting the comparatively short predictable time horizon of individual genesis events relative to the more persistent large-scale circulation anomalies underlying heatwave and precipitation extremes.

4. DISCUSSION

These results show that machine learning models provide consistent, measurable skill improvements over climatological and persistence baselines for extreme event frequency prediction across three distinct hazard types, extending prior architecture-specific demonstrations of machine learning skill for individual prediction tasks such as medium-range forecasting [4,5] and nowcasting [6,7] to a systematic, multi-hazard, multi-architecture comparison.

The reversal in relative architecture ranking between heatwaves and extreme precipitation, with the spatially aware ConvLSTM leading for heatwaves and the tree-based XGBoost leading for precipitation, suggests that architecture choice should be informed by the expected spatial scale of the governing physical mechanism rather than defaulting to a single preferred architecture across hazard types, consistent with prior observations that different geoscientific prediction tasks reward different inductive biases [15,16]. This has a practical implication for operational extreme-event early-warning systems that may need to deploy different model architectures for different hazards rather than a single unified model.

The comparatively low and rapidly lead-time-degrading skill for tropical cyclone genesis frequency is consistent with the shorter intrinsic predictability of individual genesis events relative to the more persistent large-scale patterns underlying heatwave and precipitation extremes, and indicates that machine learning gains for this hazard type may depend more on incorporating genesis-specific dynamical predictors, such as vertical wind shear and tropical wave activity, than on architecture choice alone [17].

Several limitations should be considered. First, the study used a fixed 1979-2015/2016-2023 temporal split rather than multiple independent train/test partitions, and absolute skill estimates could differ somewhat under an alternative split, though the relative ranking of architectures by event type would be expected to be reasonably robust to this choice. Second, predictor fields were drawn exclusively from ERA5 reanalysis; skill in an operational forecasting context using forecast-model initial conditions rather than reanalysis may differ, particularly at longer lead times. Third, this analysis considered only three event types and three geographic domains, and skill patterns identified here should not be assumed to generalise directly to other regions or hazard types, such as extratropical windstorms or drought, without dedicated evaluation.

5. CONCLUSION

A systematic comparison of random forest, gradient-boosted decision tree, and hybrid convolutional-LSTM models for predicting weekly frequency of heatwaves, extreme precipitation, and tropical cyclone genesis shows that all three machine learning architectures outperform climatological and persistence baselines across all event types and lead times tested, that the convolutional-LSTM architecture holds a skill advantage for the more spatially coherent heatwave signal while gradient boosting holds a modest advantage for more localized extreme precipitation, and that skill for tropical cyclone genesis remains comparatively low and short-lived across all architectures. These findings suggest that operational extreme-event early warning could benefit from architecture selection informed by the spatial scale of the governing hazard mechanism rather than a single default model.

Acknowledgements

The authors thank the reanalysis data infrastructure team at the Shannon University Centre for Ecology and Biodiversity for ERA5 data access and processing support, and colleagues at the Iberian University Center for Environmental Sustainability for helpful discussion of the model evaluation framework. We also thank the maintainers of the IBTrACS best track dataset used in this study.

Funding

This work was supported by internal research funds from the Shannon University Centre for Ecology and Biodiversity, the Iberian University Center for Environmental Sustainability, and the Baltic Institute of Renewable Energy. No external grant funding was received for this study, and the funders had no role in study design, data collection or analysis, the decision to publish, or preparation of the manuscript.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

L.L. conceived and supervised the study, developed the ConvLSTM architecture, and led the writing of the manuscript. O.O. curated and processed the ERA5 predictor fields and performed the tree-based model training and cross-validation. A.T. performed the skill scoring, lead-time evaluation, and feature importance analysis. All authors contributed to writing and approved the final manuscript.

Ethics Statement

This study used only publicly available ERA5 reanalysis output and the publicly available IBTrACS best track dataset, and did not involve human participants, animal subjects, or field data collection requiring ethical review.

Data Availability

ERA5 reanalysis data used in this study is publicly available through the Copernicus Climate Data Store, and IBTrACS

best track data is publicly available through NOAA National Centers for Environmental Information. Processed training datasets, trained model weights, and analysis scripts supporting the findings of this study are available from the corresponding author on reasonable request.

REFERENCES

1. Boukabara SA, Krasnopolsky V, Stewart JQ, Maddy ES, Shahroudi N, Hoffman RN (2019) Outlook for exploiting artificial intelligence in the Earth and environmental sciences. *Bulletin of the American Meteorological Society* 100(12):2473-2490. <https://doi.org/10.1175/BAMS-D-18-0031.1>
2. McGovern A, Elmore KL, Gagne DJ, Haupt SE, Karstens CD, Lagerquist R, Smith T, Williams JK (2017) Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society* 98(10):2073-2090. <https://doi.org/10.1175/BAMS-D-16-0123.1>
3. Chantry M, Christensen H, Dueben P, Palmer T (2021) Opportunities and challenges for machine learning in weather and climate modelling: hard, medium and soft AI. *Philosophical Transactions of the Royal Society A* 379(2194):20200083. <https://doi.org/10.1098/rsta.2020.0083>
4. Bi K, Xie L, Zhang H, Chen X, Gu X, Tian Q (2023) Accurate medium-range global weather forecasting with 3D neural networks. *Nature* 619(7970):533-538. <https://doi.org/10.1038/s41586-023-06185-3>
5. Lam R, Sanchez-Gonzalez A, Willson M, Wirsberger P, Fortunato M, Alet F, Ravuri S, Ewalds T, Eaton-Rosen Z, Hu W, Merose A, Hoyer S, Holland G, Vinyals O, Stott J, Pritzel A, Mohamed S, Battaglia P (2023) Learning skillful medium-range global weather forecasting. *Science* 382(6677):1416-1421. <https://doi.org/10.1126/science.adi2336>
6. Ravuri S, Lenc K, Willson M, Kangin D, Lam R, Mirowski P, Fitzsimons M, Athanassiadou M, Kashem S, Madge S, Prudden R, Mandhane A, Clark A, Brock A, Simonyan K, Hadsell R, Robinson N, Clancy E, Arribas A, Mohamed S (2021) Skillful precipitation nowcasting using deep generative models of radar. *Nature* 597(7878):672-677. <https://doi.org/10.1038/s41586-021-03854-z>
7. Espeholt L, Agrawal S, Sønderby C, Kumar M, Heek J, Bromberg C, Gaze C, Carver R, Andrychowicz M, Hickey J, Bell A, Kalchbrenner N (2022) Deep learning for twelve hour precipitation forecasts. *Nature Communications* 13:5145. <https://doi.org/10.1038/s41467-022-32483-x>
8. Ham YG, Kim JH, Luo JJ (2019) Deep learning for multi-year ENSO forecasts. *Nature* 573(7775):568-572. <https://doi.org/10.1038/s41586-019-1559-7>
9. Prabhat, Kashinath K, Mudigonda M, Kim S, Kapp-Schwoerer L, Graubner A, Karaismailoglu E, von Kleist L, Kurth T, Greiner A, Mahesh A, Yang K, Lewis C, Chen J, Lou A, Chandran S, Toms B, Chapman W, Dagon K, Shields CA, O'Brien T, Wehner M, Collins W (2021) ClimateNet: an expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development* 14(1):107-124. <https://doi.org/10.5194/gmd-14-107-2021>
10. Chattopadhyay A, Hassanzadeh P, Pasha S (2020) Predicting clustered weather patterns: a test case for applications of convolutional neural networks to spatio-temporal climate data. *Scientific Reports* 10:1317. <https://doi.org/10.1038/s41598-020-57897-9>
11. Gagne DJ, Haupt SE, Nychka DW, Thompson G (2019) Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review* 147(8):2827-2845. <https://doi.org/10.1175/MWR-D-18-0316.1>
12. Herman GR, Schumacher RS (2018) Money doesn't grow on trees, but forecasts do: forecasting extreme precipitation with random forests. *Monthly Weather Review* 146(5):1571-1600. <https://doi.org/10.1175/MWR-D-17-0250.1>
13. Schultz MG, Betancourt C, Gong B, Kleinert F, Langguth M, Leufen LH, Mozaffari A, Stadler S (2021) Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A* 379(2194):20200097. <https://doi.org/10.1098/rsta.2020.0097>

14. Watson-Parris D (2021) Machine learning for weather and climate are worlds apart. *Philosophical Transactions of the Royal Society A* 379(2194):20200098. <https://doi.org/10.1098/rsta.2020.0098>
15. Toms BA, Barnes EA, Ebert-Uphoff I (2020) Physically interpretable neural networks for the geosciences: applications to Earth system variability. *Journal of Advances in Modeling Earth Systems* 12(9):e2019MS002002. <https://doi.org/10.1029/2019MS002002>
16. Barnes EA, Hurrell JW, Ebert-Uphoff I, Anderson C, Anderson D (2019) Viewing forced climate patterns through an AI lens. *Geophysical Research Letters* 46(22):13389-13398. <https://doi.org/10.1029/2019GL084944>
17. Giffard-Roisin S, Yang M, Charpiat G, Kumler Bonfanti C, Kegl B, Monteleoni C (2020) Tropical cyclone track forecasting using fused deep learning from aligned reanalysis data. *Frontiers in Big Data* 3:1. <https://doi.org/10.3389/fdata.2020.00001>
18. Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat (2019) Deep learning and process understanding for data-driven Earth system science. *Nature* 566(7743):195-204. <https://doi.org/10.1038/s41586-019-0912-1>
19. Rasp S, Pritchard MS, Gentile P (2018) Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences* 115(39):9684-9689. <https://doi.org/10.1073/pnas.1810286115>
20. Rasp S, Dueben PD, Scher S, Weyn JA, Mouatadid S, Thuerey N (2020) WeatherBench: a benchmark dataset for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems* 12(11):e2020MS002203. <https://doi.org/10.1029/2020MS002203>
21. Weyn JA, Durran DR, Caruana R (2019) Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems* 11(8):2680-2693. <https://doi.org/10.1029/2019MS001705>
22. Kurth T, Treichler S, Romero J, Mudigonda M, Luehr N, Phillips E, Mahesh A, Matheson M, Deslippe J, Fatica M, Houston M, Prabhat (2018) Exascale deep learning for climate analytics. *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC18)*, pp. 649-660. <https://doi.org/10.1109/SC.2018.00054>
23. Huntingford C, Jeffers ES, Bonsall MB, Christensen HM, Lees T, Yang H (2019) Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters* 14(12):124007. <https://doi.org/10.1088/1748-9326/ab4e55>
24. Scher S (2018) Toward data-driven weather and climate forecasting: approximating a simple general circulation model with deep learning. *Geophysical Research Letters* 45(22):12616-12622. <https://doi.org/10.1029/2018GL080704>
25. Davenport FV, Diffenbaugh NS (2021) Using machine learning to analyze physical causes of climate change: a case study of U.S. Midwest extreme precipitation. *Geophysical Research Letters* 48(15):e2021GL093787. <https://doi.org/10.1029/2021GL093787>
26. Vandal T, Kodra E, Ganguly S, Michaelis A, Nemani R, Ganguly AR (2017) DeepSD: generating high resolution climate change projections through single image super-resolution. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1663-1672. <https://doi.org/10.1145/3097983.3098004>
27. Karpatne A, Ebert-Uphoff I, Ravela S, Babaie HA, Kumar V (2019) Machine learning for the geosciences: challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering* 31(8):1544-1554. <https://doi.org/10.1109/TKDE.2018.2861006>
28. Camps-Valls G, Tuia D, Zhu XX, Reichstein M (eds) (2021) *Deep learning for the Earth Sciences: a comprehensive approach to remote sensing, climate science and geosciences*. Wiley. <https://doi.org/10.1002/9781119646181>