Near real-time identification of extreme events for weather index insurance using machine learning algorithms

EGU General Assembly 2021

vEGU21: Gather Online | 19th-30th April 2021

Luigi Cesarini

luigi.cesarini@iusspavia.it

CARISMA Centre

IUSS Scuola Universitaria Superiore Pavia

The SMART project

- Our consortium:

IUSS Scuola Universitaria Superiore Pavia

UNIVERSITY OF EXETER

Fundación REDDOM RURAL ECONOMIC DEVELOPMENT DOMINICANA

- Thematic areas of the call addressed by our proposal:
  - Machine Learning and Big Data for Disaster Risk Financing.
  - Disaster Risk Financing Mechanisms to Manage Food Insecurity.

Project Objectives

- Develop an innovative framework for the design of parametric triggers for weather index insurance, based on machine learning methods.

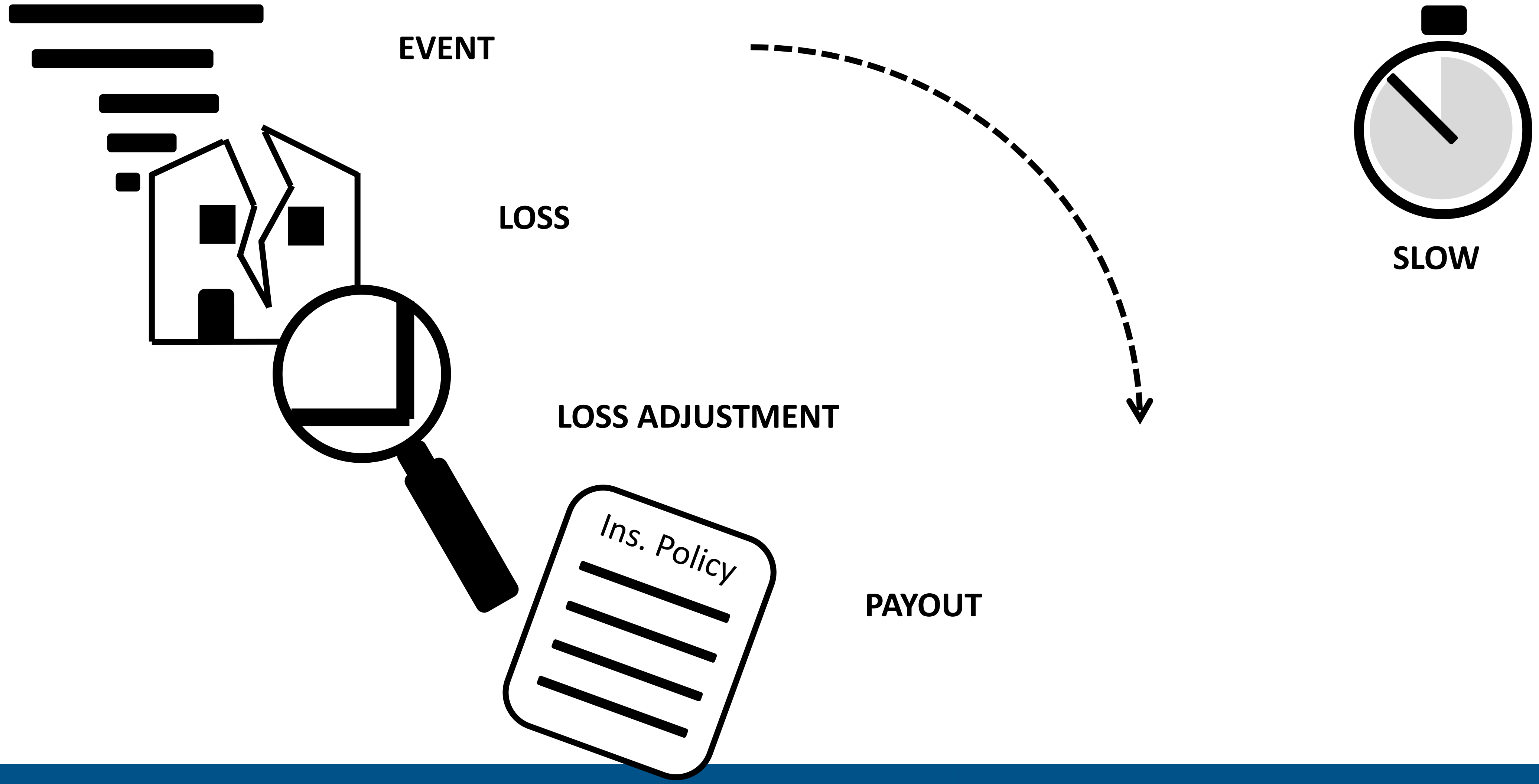
- Demonstrate the framework and its pathway to operationalization in a pilot study of multiple hazards in the Dominican Republic.

- Special focus on the agricultural sector, which tends to be severely affected by natural hazards.

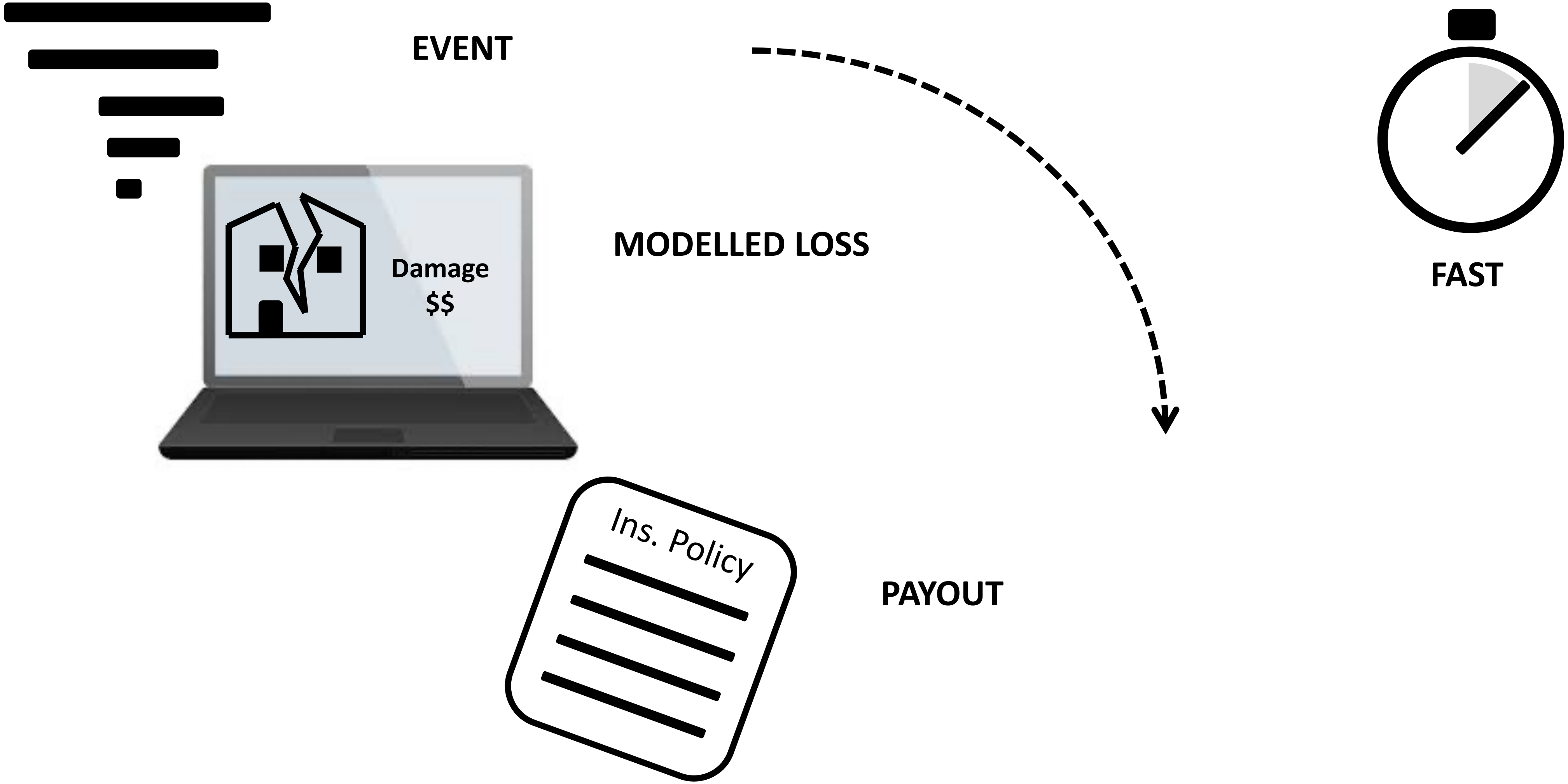
Index Insurance

- Innovative insurance instrument
- Payouts are based on an index (e.g. rainfall level) rather than on assessed losses
- Clients get paid if index exceeds a value pre-determined in the insurance policy
- The index measures deviations from the normal level of parameters such as rainfall, temperature, wind speed, crop yield and livestock mortality rates.

Traditional Insurance



Index Insurance



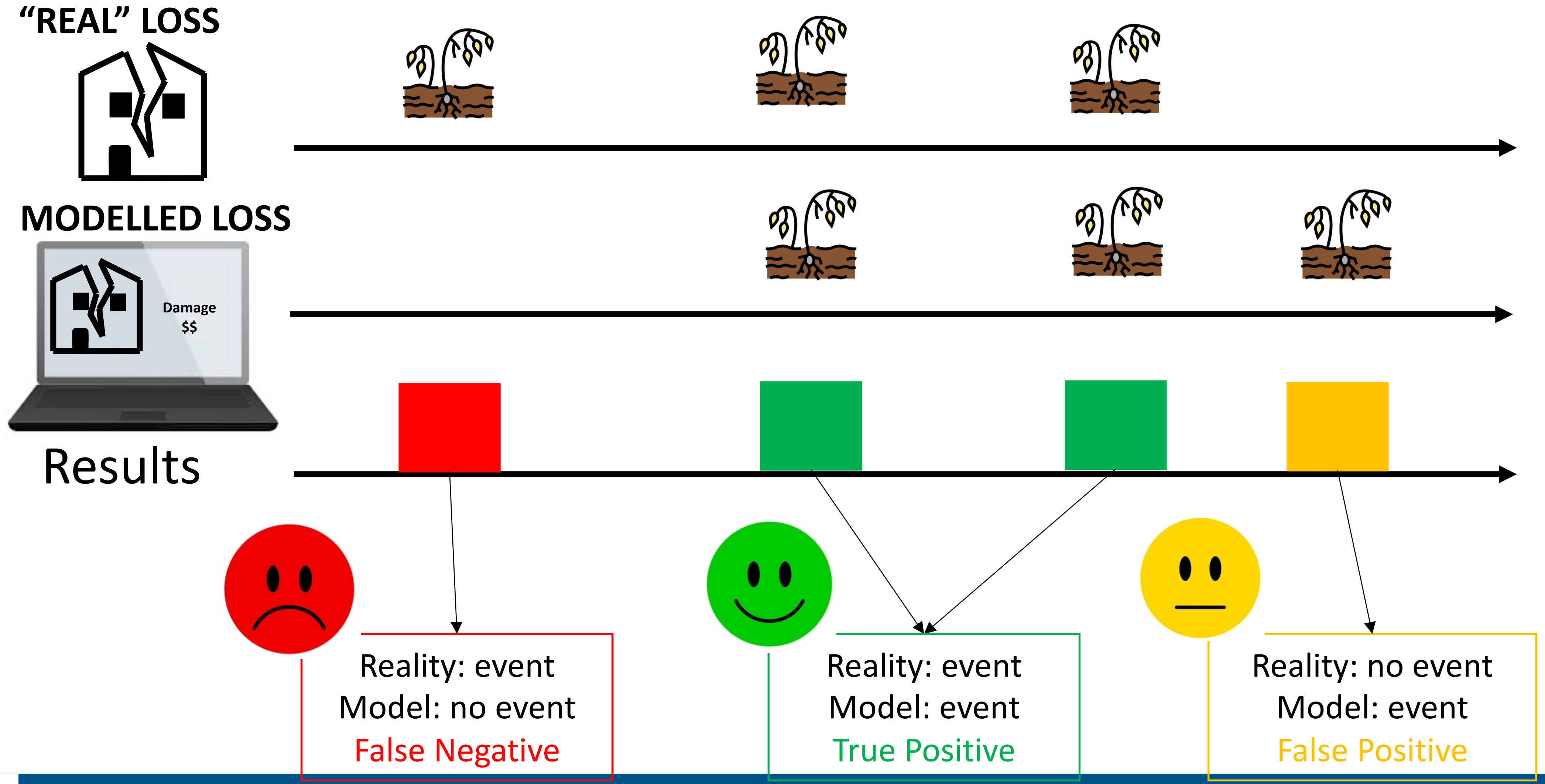
Index insurance: pros and cons

- Payouts are based on observed environmental variables (indices)
- Fast funding after disaster
- Low administrative costs
- Index values cannot be manipulated

- Modelled index may not always reflect actual losses, i.e. basis risk: users may suffer losses and receive no payout, or vice-versa

- Key challenge – minimizing basis risk

Basis Risk



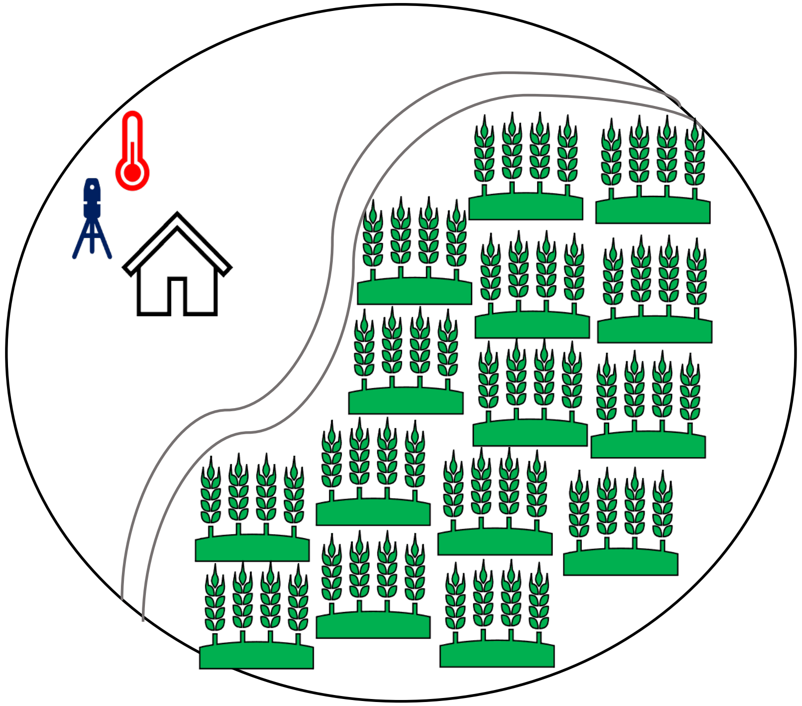
“REAL” LOSS
MODELLED LOSS
Damage
$$
Results
Reality: event
Model: no event
False Negative
Reality: event
Model: event
True Positive
Reality: no event
Model: event
False Positive

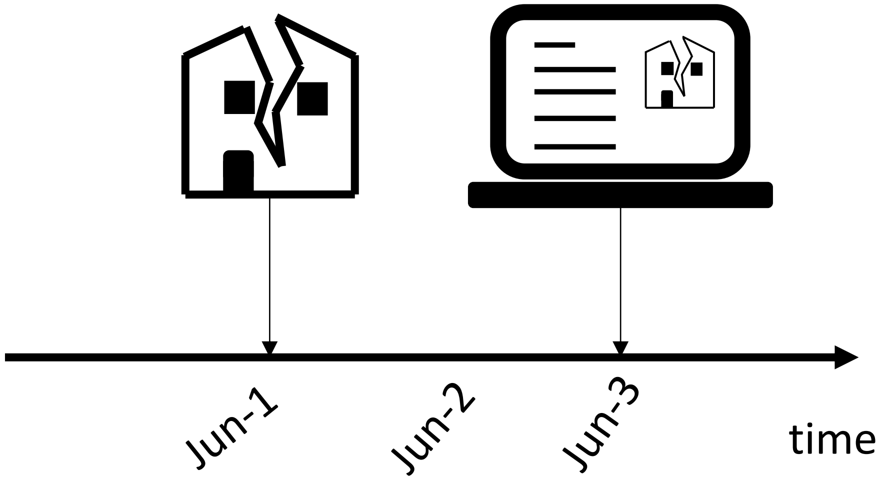
Basis Risk

Basis risk -> mismatch between payouts and occurrence of losses

Spatial basis risk: when the station responsible for determining the index is far from the insured location

Temporal basis risk: when the model identifies the event, but too early or too late





Case Study: Dominican Republic

Task: Identification of Flood and Drought events between 2000-2019

Input Data

Rainfall Dataset

Dataset	Type	Resolution	Time Span	Latency
CCS	Satellite	0.04°	Jan 2003-Today	6 Hours
CHIRPS	Satellite-Gauge	0.05°	Jan 1981-Today	3 Weeks
CMORPH	Satellite-Gauge	0.07°	Jan 1998-Today	14 Days
GSMaP	Satellite	0.1°	Mar 2000-Today	12 Hours
IMERG	Satellite	0.1°	Jun 2000-Today	12 Hours
PERSIANN	Satellite	0.25°	Mar 2000-Today	48 Hours

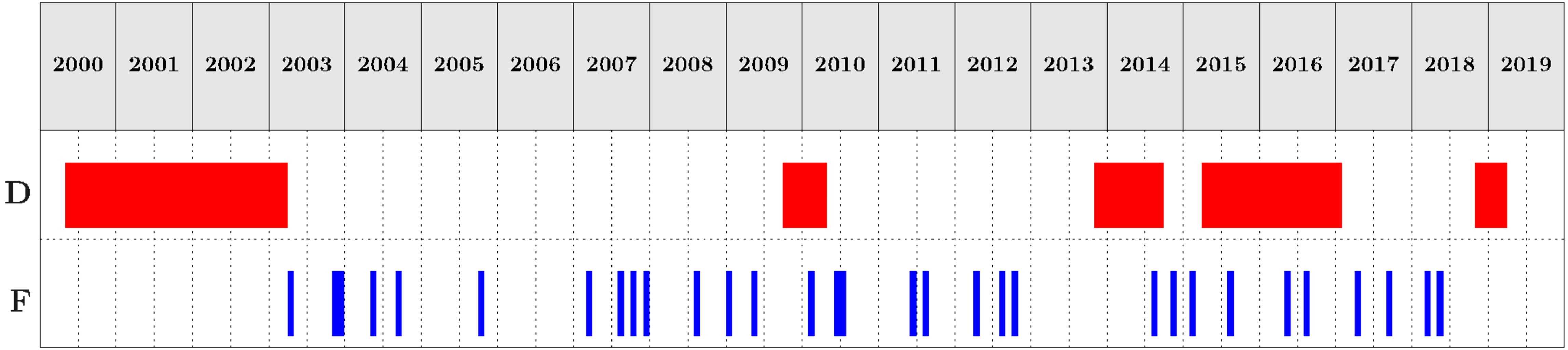
Soil Moisture Dataset

Dataset	Type	Resolution	Time Span	Latency
ERA5	Reanalsysis	0.25°	Jan 1979-Today	5 Days

Case Study: Dominican Republic

Target (output) Dataset:

Catalogue of Historical event (2000- 2019)
Sources:
EM-DAT, DesInventar, GLIDE, DFO, FLoodList



Methodology proposed

Two machine learning algorithm:

- Neural Network (NN)
- Support Vector Machine (SVM)

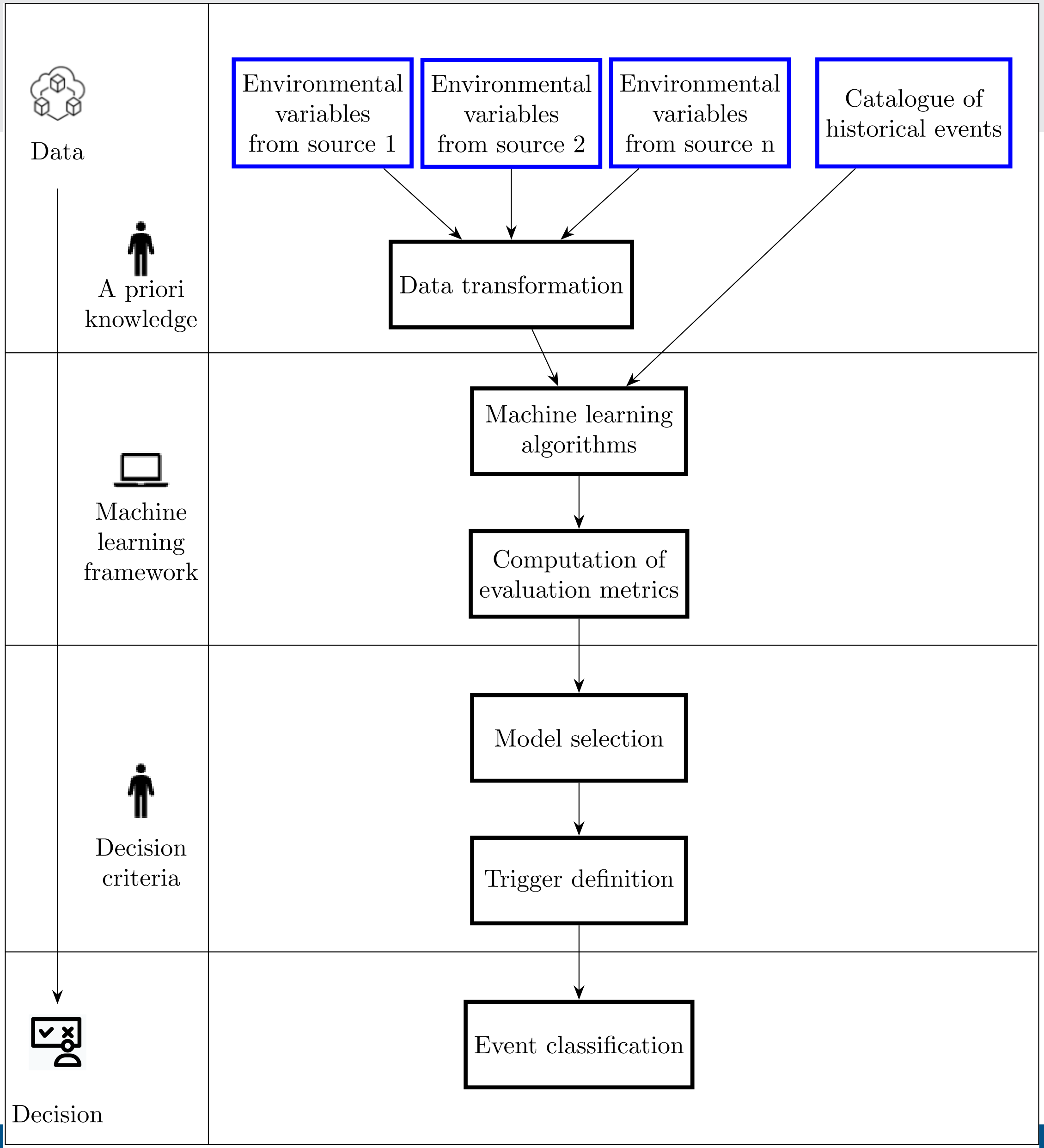
Data Transformation from rainfall:

- Potential damage
- SPI

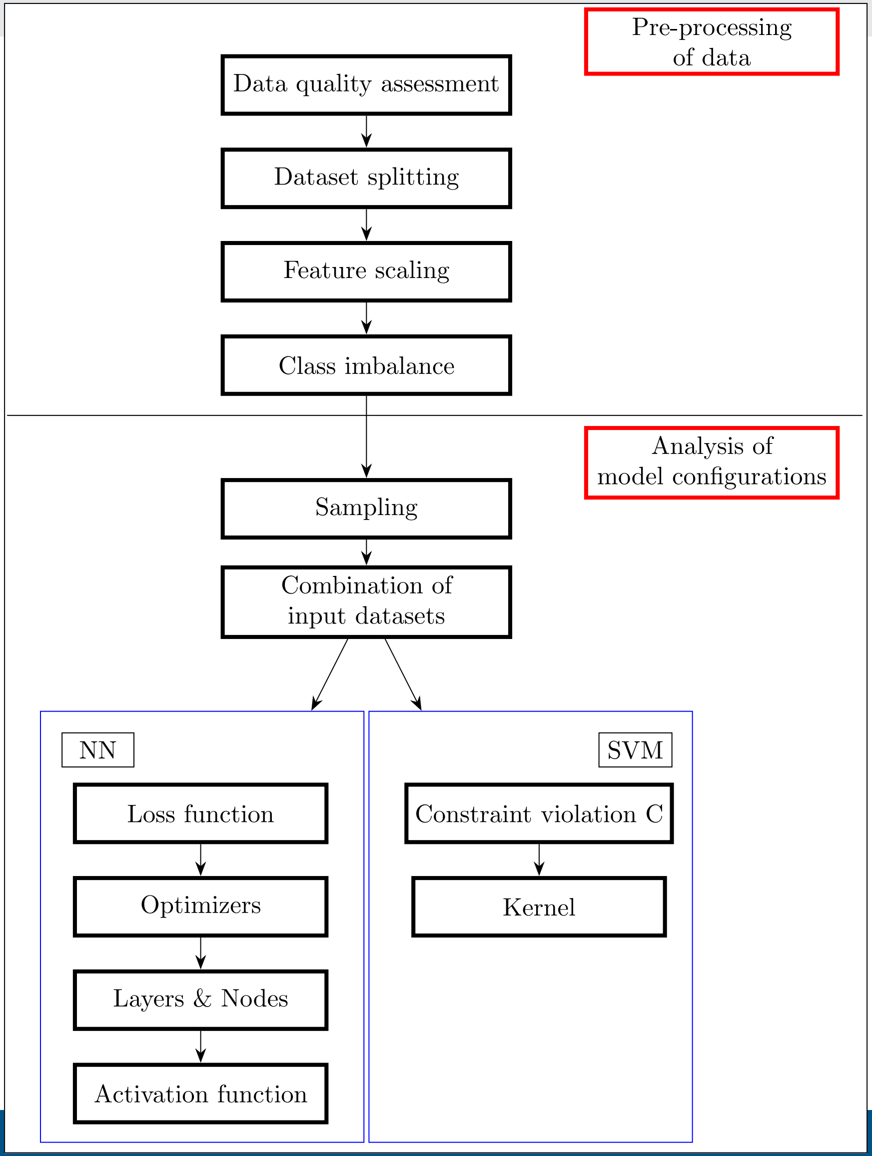
Evaluation through confusion matrix:

- F1 Score as main metrics
- Precision-Recall curve more suitable in an imbalanced data problem

The machine learning algorithms box has its own framework.



Machine Learning framework

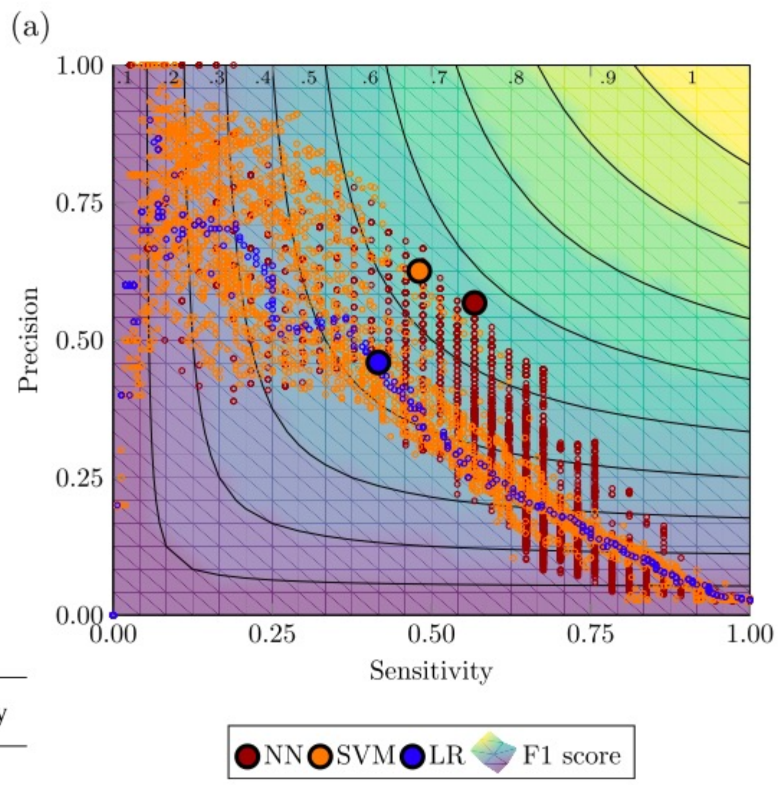


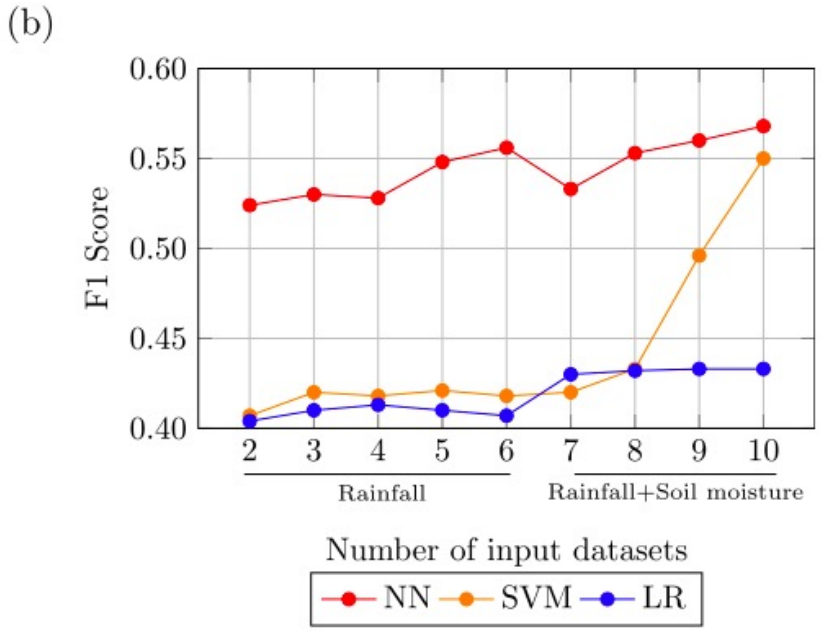
Model	Parameter	Flood	Drought
NN	Input dataset combinations	61 combinations of environmental variables	61 combinations of environmental variables 4 SPI (1,3,6,12)
	Sampling	Unweighted Class Weight Over-sampling SMOTE	Unweighted Class Weight
	Loss	Binary Cross Entropy	Binary Cross Entropy
	Optimizer	ADAM	ADAM
	Number of layers & nodes	Layers: $[1;9]$ Nodes: $2^{nl+1}:2^{nl+9}(*)$	Layers: $[1;9]$ Nodes $2^{nl+1}:2^{nl+9}(*)$
	Activations	ReLu Tanh	ReLu Tanh
Number of Configurations		4392	8784
SVM	Input dataset combinations	67 combinations of environmental variables	67 combinations of environmental variables 4 SPI (1,3,6,12)
	Sampling technique	Unweighted Class Weight Over-sampling SMOTE	Unweighted Class Weight
	C-Regularization parameter	$C=(0.1,1,10,100,500)$	$C=(0.1,1,10,100,500)$
	Kernel Function	Linear Polynomial Radial Basis	Linear Polynomial Radial Basis
Number of Configurations		4020	8040

$(*)$: nl: number of layers

Results: Flood (1)

- NN and SVM outperforms logistic regression (LR), used in this work as a benchmark, according to F1 score (a).
- ML models improve their performance with increasing number of input dataset (b).

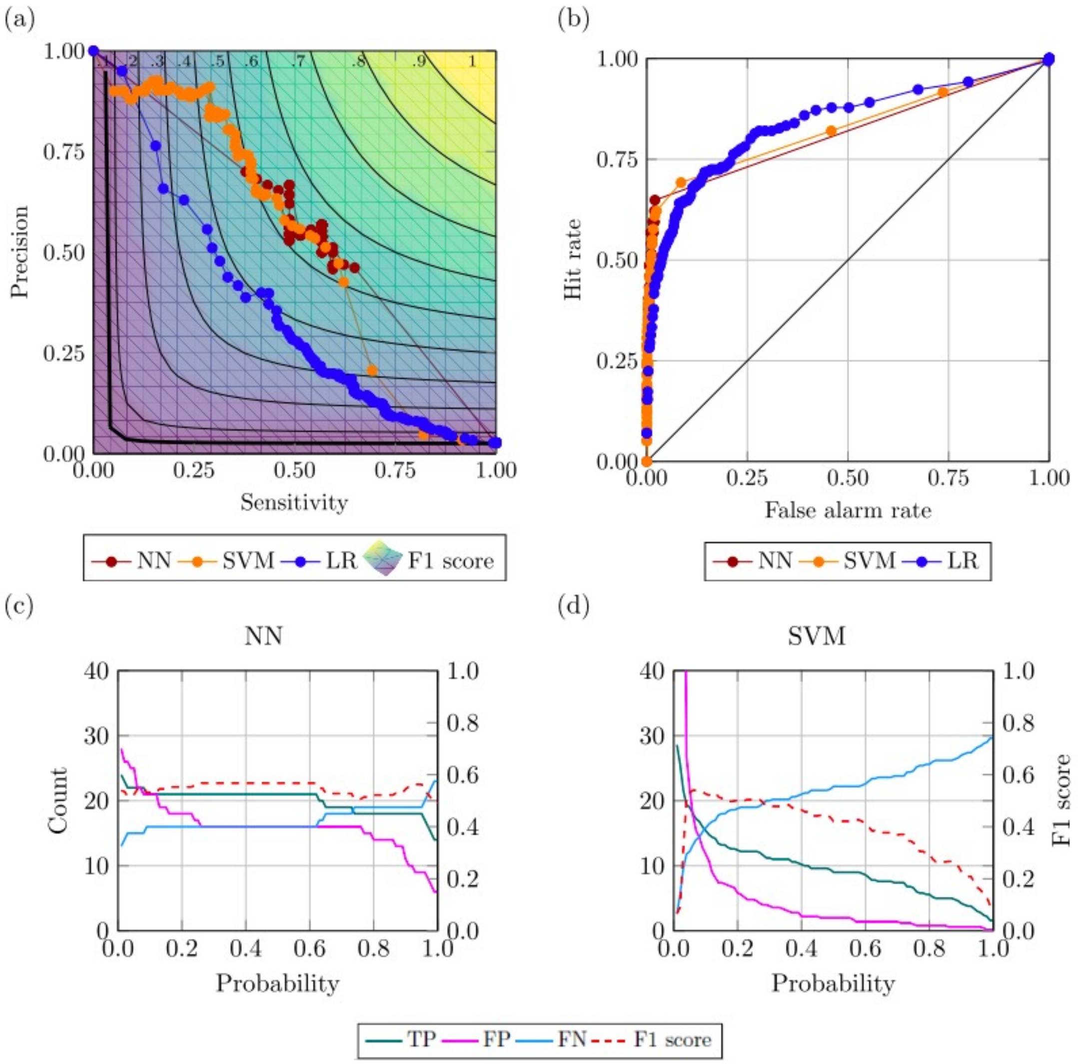




Method	Precision	Sensitivity	Specificity	F1 Score	Accuracy
NN	0.57	0.57	0.99	0.57	0.98
SVM	0.63	0.49	0.99	0.55	0.98
LR	0.46	0.42	0.99	0.43	0.97

Results: Flood (2)

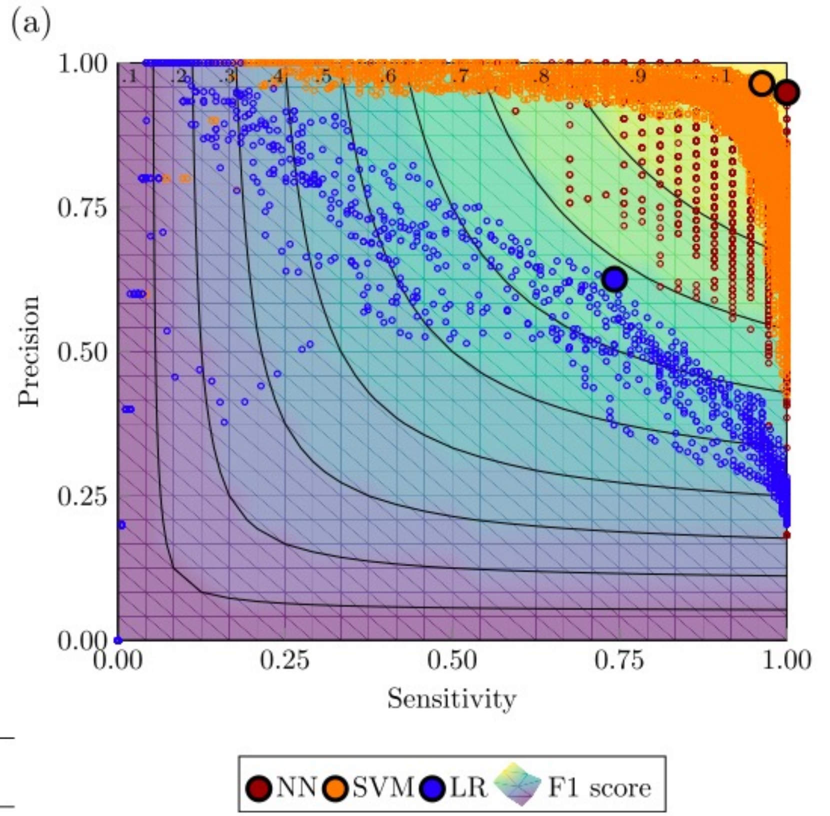
- In (a) and (b), precision-sensitivity (PS) and ROC curve for the best-performing configuration, which we defined as the one with the highest AUC under the PS curve.
- (c) and (d) exhibit the behaviour of the predictions when changing the probability threshold.
- The predictions steadiness of the NN in (c) might be linked to the confidence calibration of modern neural network (Guo et al., 2017).

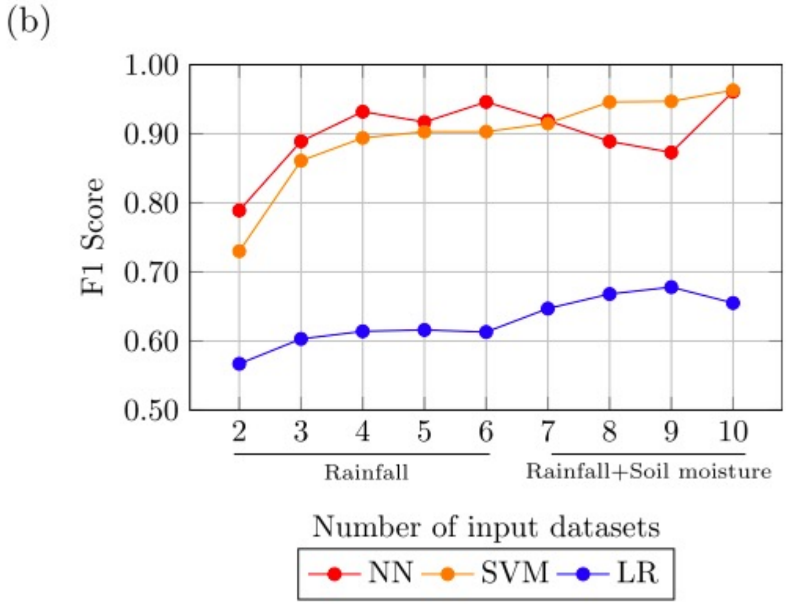


Results: Drought (1)

- Each week of drought was counted as an event for the binary classification. This might explain the high values over all the metrics.

- Regardless, the machine learning models outperform LR even more evidently.



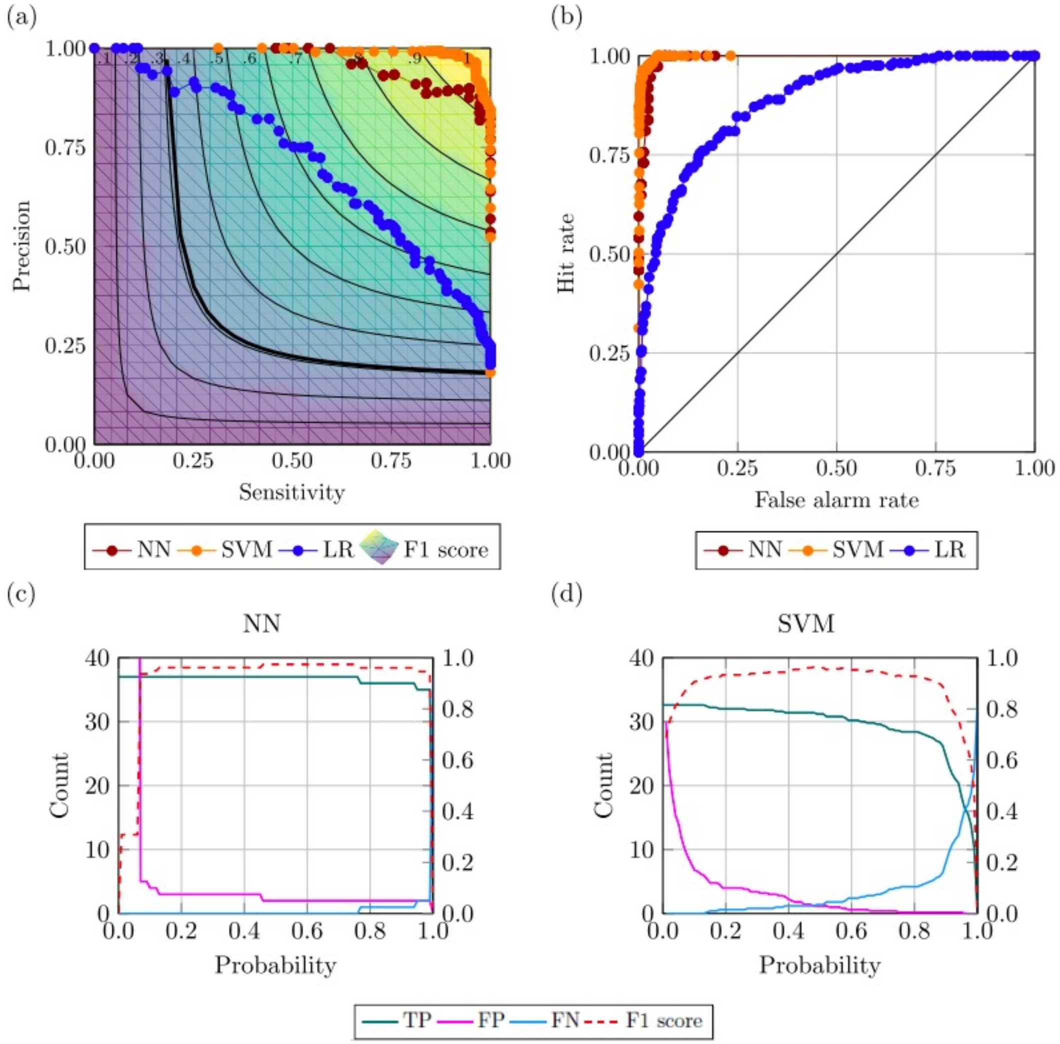


Method	Precision	Sensitivity	Specificity	F1 Score	Accuracy
NN	0.95	1.00	0.99	0.97	0.99
SVM	0.96	0.96	0.99	0.96	0.98
LR	0.63	0.74	0.89	0.68	0.85

Results: Drought (2)

- Similarly to the flood case, the PS and ROC curve confirm that ML models provide an higher quality of the prediction.

- In (c) we encounter the same suspicious behaviour. Problems regarding the confidence of the probability estimates is an open research question that which goes beyond the aim of this work and warrants further research



Conclusion

- Using ML we are able to reduce basis risk w.r.t traditional method such as Logistic Regression
- Importance of data enhancing technique and selection of adequate evaluation criteria when dealing with extreme events
- ML model better equipped to handle increase number of information
- The capability of these algorithms to rely on global data that are disentangled from the resources of a given territory, both from the point of view of climate data (e.g., lack of rain-gauge network) and from the point of view of information about past natural disasters, is an appealing feature of this work that would be a first step towards increasing the reliability of weather index insurance program.

Cesarini, L., Figueiredo, R., Monteleone, B., and Martina, M. L. V.: The potential of big data and machine learning for weather index insurance, Nat. Hazards Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/nhess-2020-220, in review, 2020.

Thank you for your attention!

luigi.cesarini@iusspavia.it

Carisma Centre Website