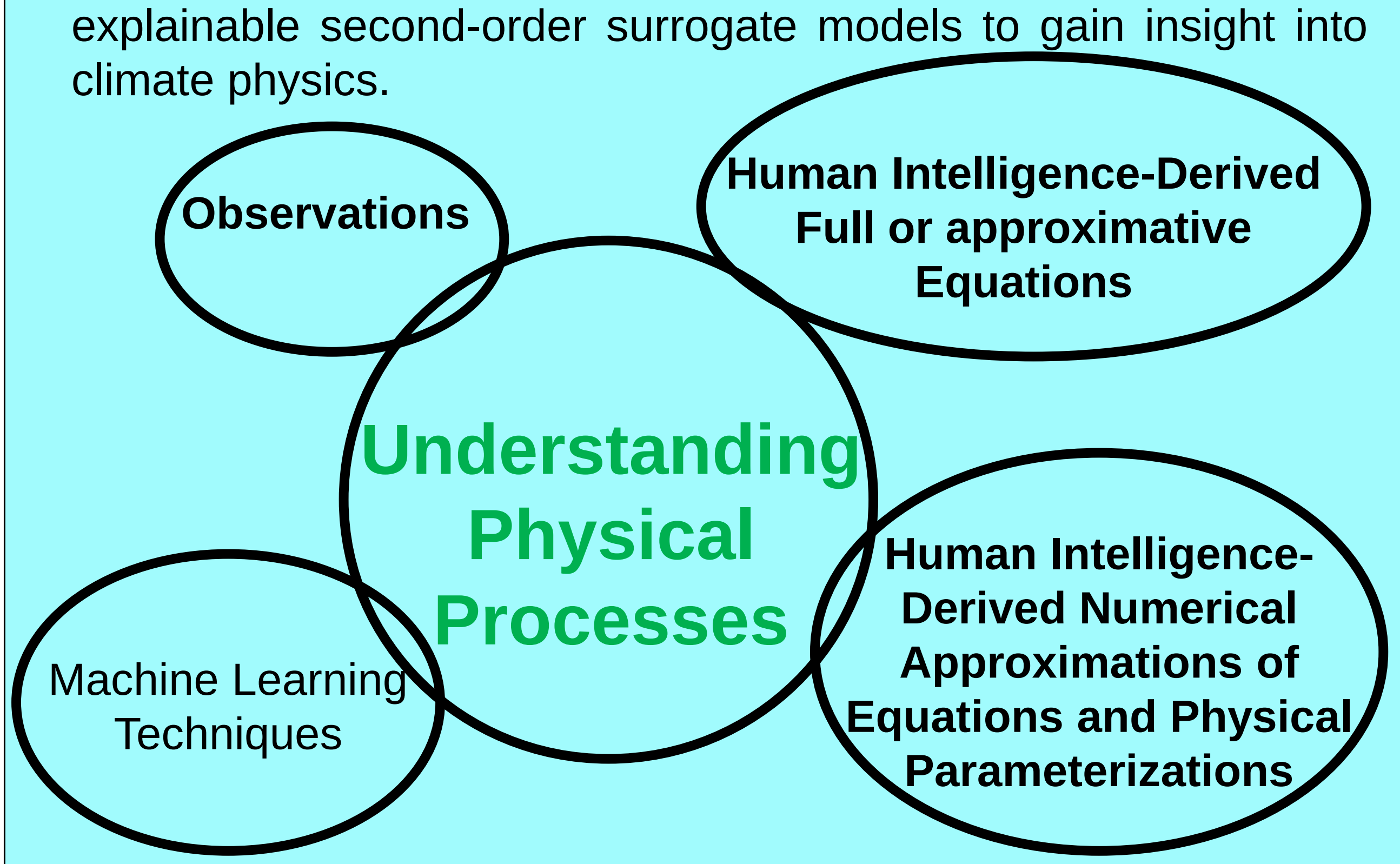
Using Decision Trees for **Data-Driven Process Studies,** Conditional Bias Correction, and Model Physics Improvements

Augustin Vintzileos[1], Benjamin Cash[2], and Jim Kinter[2]

[1]University of Maryland – ESSIC [2] George Mason University - COLA

Problem: Investigating the physics of climate is hampered by a limited number of observations and the inability to conduct A/B (perturbation) experiments on the natural system. Computer simulations resolving the equations of motion provide the first-order surrogate solution. However, numerical models are imperfect, especially when representing subgrid-scale phenomena (model physics).

Solution: Combine observational and model-generated data, human expertise, and machine-learning techniques to create explainable second-order surrogate models to gain insight into climate physics.

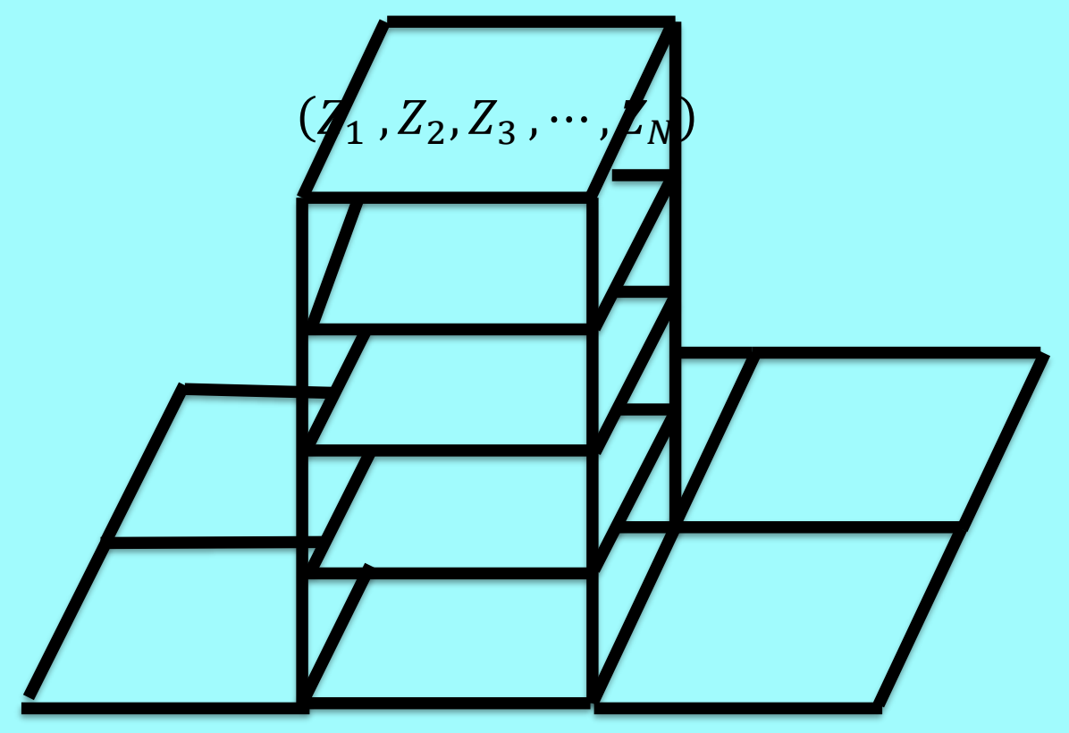

Observations
Human Intelligence-Derived Full or approximative Equations
Understanding Physical Processes
Human Intelligence-Derived Numerical Approximations of Equations and Physical Parameterizations
Machine Learning Techniques

Methodology: In what follows, we consider target data to be tendencies of a variable or errors in predicting this variable. We aggregate the target data from all grid points and every time step into an unconditional distribution. This aggregation is based on the postulate that physics is geographically and temporally agnostic. A postulate which led to fruitful results when applied by Hewson and Pilosu (2020) at ECMWF.

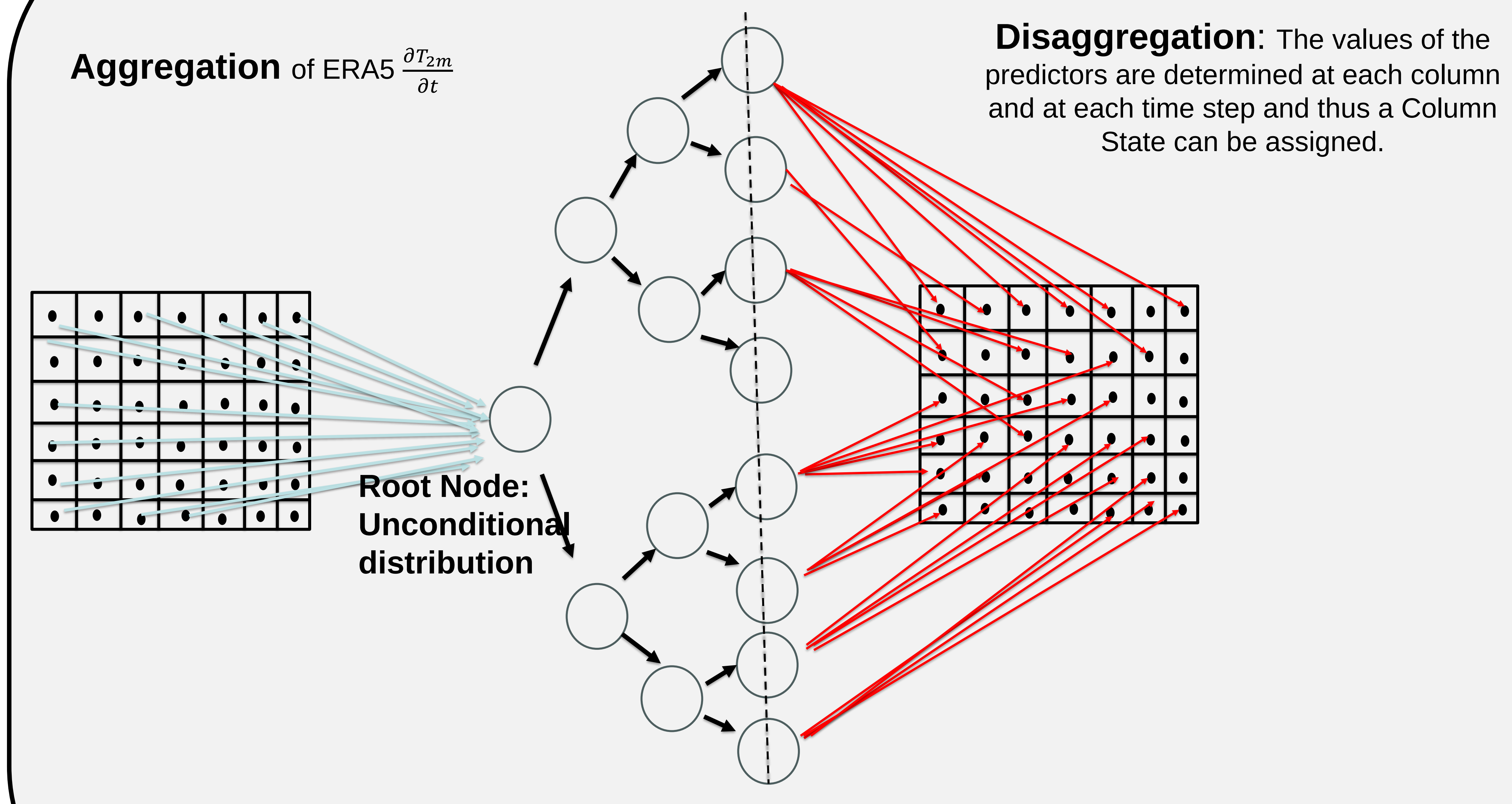
The human expert chooses a subset of predictors (features) from the model column that are relevant to the target variable thus reducing the problem's dimensionality. The unconditional distribution of the target variable is then progressively split into nodes based on optimization criteria, e.g., entropy and variance. Splitting continues up to the leaves of the decision tree, each one of which associates the state of a model column to the distribution of the target variable.

The decision tree grows to a fully explainable ML surrogate model, which can be disaggregated at each grid point and time step of a model simulation. Discrepancies between the numerical model and the decision tree provide evidence of the importance of the decision chains and, thus, of physical processes.

Physical parameterizations act on each of the model's columns, but their formulation does not depend on the grid point's geographical location or the simulation's time step (**physics is geographically and temporally agnostic**).



Decision tree induction by maximizing the variance of the split (regression tree)


Aggregation of ERA5 $\frac{\partial T_{2m}}{\partial t}$

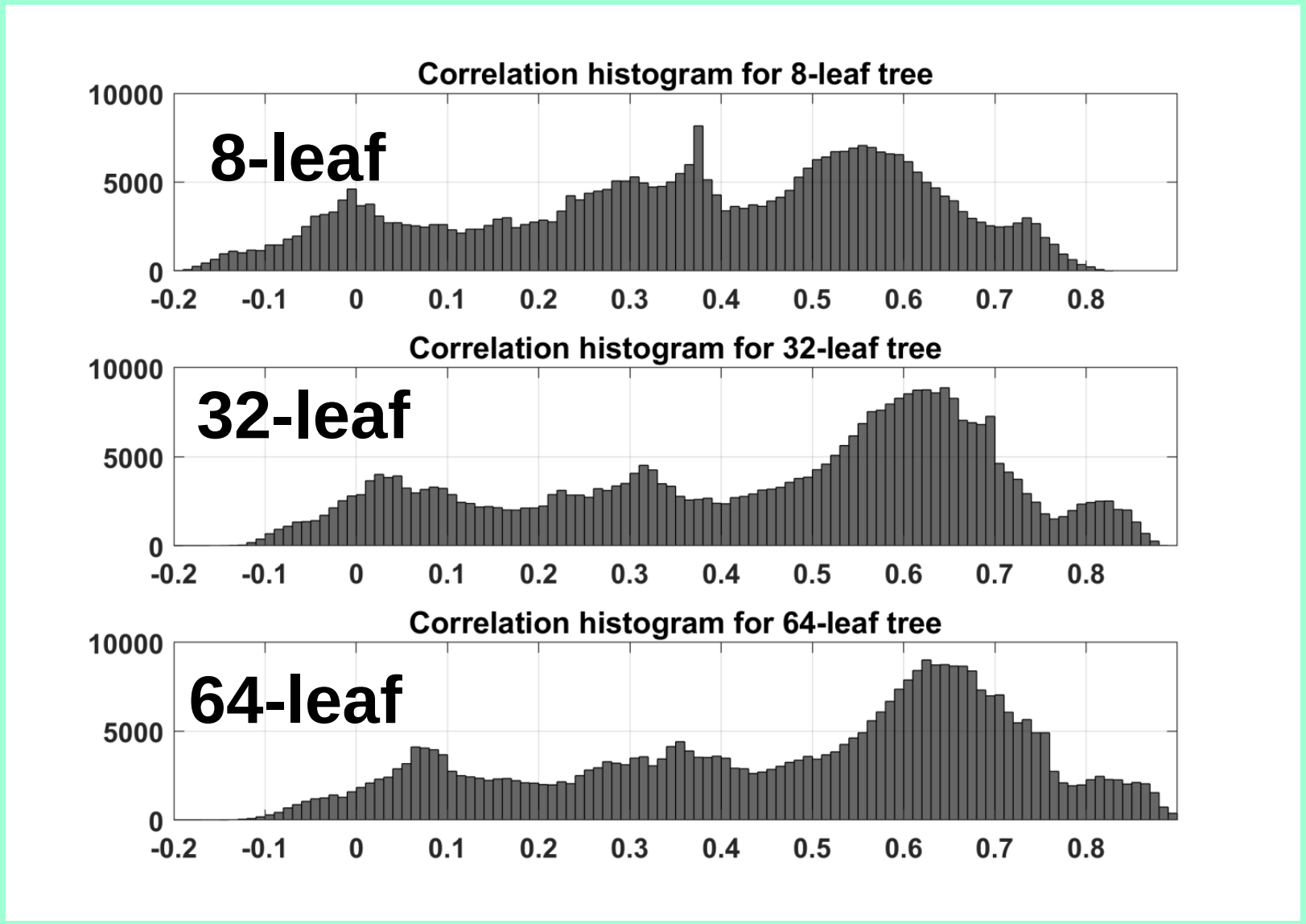
Disaggregation: The values of the predictors are determined at each column and at each time step and thus a Column State can be assigned.

Root Node: Unconditional distribution

Tree Leaves or Column States. Each CS contains a distribution of the target variable.

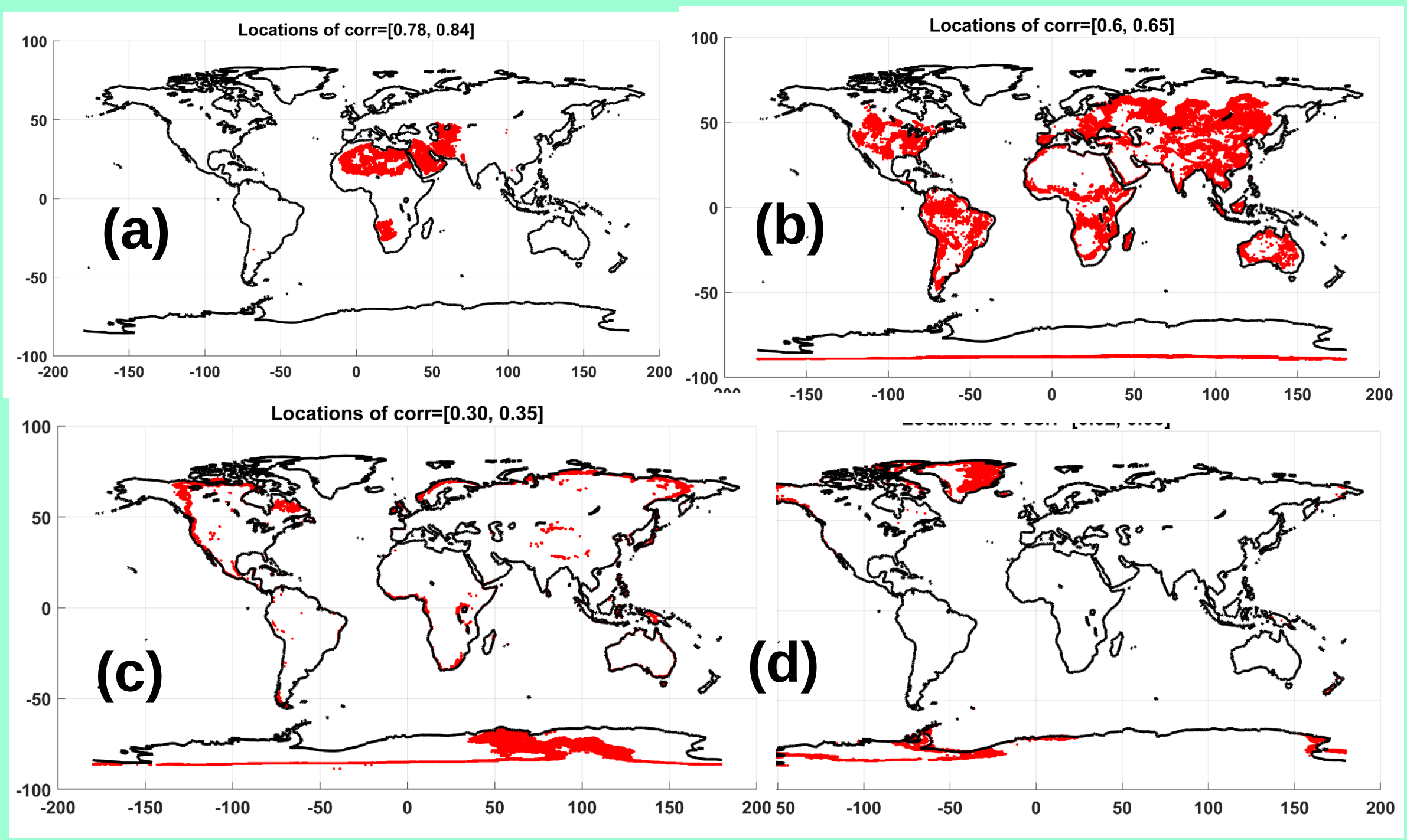
Example: 2-meter temperature tendencies from ERA5

- We use hourly 2-meter temperature tendencies on land grid points from 2006.
- We choose the following predictors: T2M, SD, SH, LH, BLH, TCC, ORO, SWVL, and TSOIL.
- We train three decision trees with 8, 32, and 64 leaves.
- We train a fourth 64-leaf tree with the additional variable DSN (snow depth)
- We compute the mean temperature tendency of the Column State for each grid point and at each hourly time step.
- We evaluate the trees by comparing (correlation) the tendencies from ERA5 and the mean tendencies of the corresponding CS.

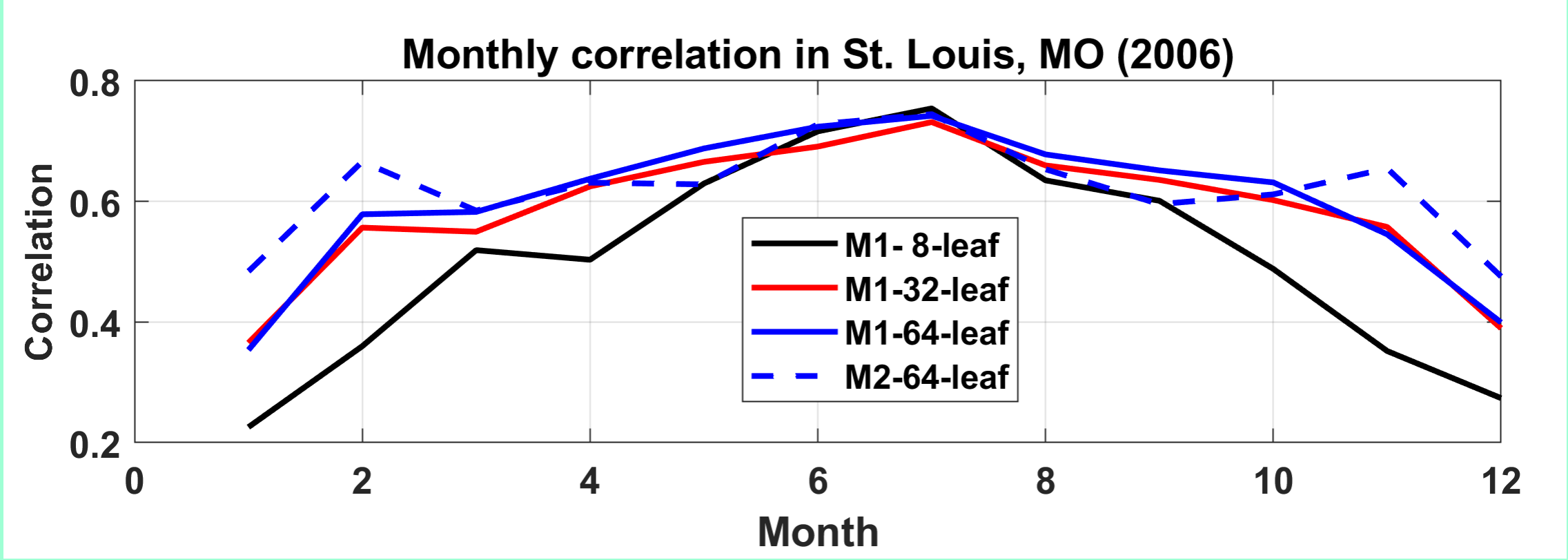


Distribution of correlation between ERA5 tendencies and mean tendencies from each Column State at all land grid points for the three decision trees.

These distributions have four poles, with correlations ranging from very low to very high. As the tree's complexity increases, these poles shift to the right, indicating that the model improves.

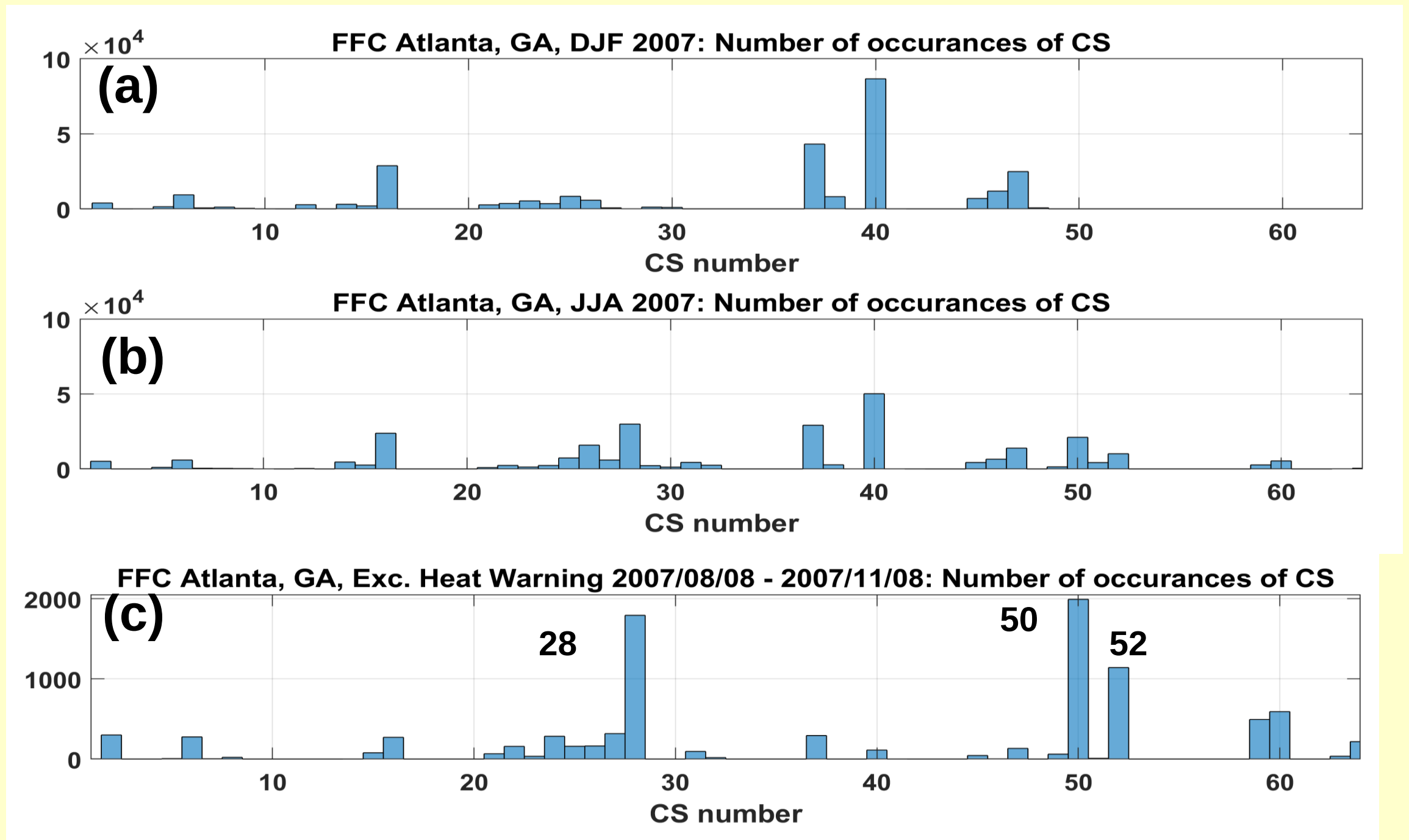

Locations of corr=[0.78, 0.84]
(a)
Locations of corr=[0.6, 0.65]
(b)
Locations of corr=[0.30, 0.35]
(c)
(d)
100
50
0
-50
-100
-200
-150
-100
-50
0
50
100
150
200

Geographical location of grid points with (a) high correlation, (b) good correlation, (c) low correlation, and (d) no correlation. Best correlations are obtained in desert areas where there is no hydrological cycle and thus the physics are simpler.



Correlations for each calendar month between ERA5 and the decision tree predictions at St. Louis, MO. During summer, all models perform well. In winter, the more complex models (32- and 64-leaf) perform better, but improvements are minimal between the 32- and 64-leaf trees. Better performance in winter is obtained by adding the snow depth to the set of predictors.

Application: Heat waves



These plots compare the number of times a given Column State appears within the FFC WFO (Atlanta, GA) in (a) winter, (b) summer, and (c) during an excessive heat wave warning issued by the WFO. We note a significant increase in the relative frequency of CS 28, 50, and 52, and decrease of CS 40. The mean value of the temperature tendency for CS-40 is -0.24 ℃ $\cdot\, hour^{-1}$

Conclusion

We introduce a novel technique that combines observational and model-generated data and human expertise using explainable ML models. The methodology aggregates a target variable into a distribution within the root node of a decision tree. Based on a set of predictors chosen using human expertise, the root node is split progressively until the leaves of the tree, or Column States, are reached. Each Column State contains a distribution of the target variable.

Using tendencies as a target variable allows studying processes, as demonstrated here, in the case of 2-meter temperature and heat waves.

Using errors in the forecast of the target variable can provide (a) a conditional bias correction system and (b) the study of the misrepresented physics.

In addition, we may be able to improve physical parameterizations by forcing single-column models with conditions compatible with selected Column States.