Generative AI for Scientific Discovery: Uncertainty and Complexity in Empirical Models

Percy Venegas[1] and Mark Kotanchek[2]

[1]King's College London

[2]Evolved Analytics LLC

Abstract

We present an account of models explanation from the philosophy of science and relate it to the practice of scientific discovery using a class of generative AI model ensembles. The key takeaway is that a complexity metric that represents explanations of the necessary length (a good tradeoff between compression and accuracy) has to be used to reduce uncertainty. We illustrate the approach with an example from catastrophe-risk and risk-hedging modeling.

Keywords: causal explanation, climate risk, generative AI, symbolic regression, complexity metrics

IA generativa para el descubrimiento científico: incertidumbre y complejidad en modelos empíricos

Resumen

Presentamos una descripción de la explicación de modelos desde la filosofía de la ciencia y la relacionamos con la práctica del descubrimiento científico utilizando una clase de conjuntos de modelos generativos de IA. La conclusión clave es que se debe utilizar una métrica de complejidad que represente explicaciones de la longitud necesaria (una buena compensación entre compresión y precisión) para reducir la incertidumbre. Ilustramos el enfoque con un ejemplo del modelo de riesgo de catástrofe y cobertura de riesgo.

Palabras clave: explicación causal, riesgo climático, IA generativa, regresión simbólica, métricas de complejidad

 doi: 10.18278/jpcs.8.1.5

将生成式人工智能用于科学发现：经验模型中的不确定性与复杂性

摘要

我们从科学哲学的视角介绍了模型解释，并使用一类生成式人工智能模型集合，将其与科学发现的实践联系起来。关键要点在于，必须使用一个能解释必要长度（即压缩与准确性之间的良好权衡）的复杂性指标，以减少不确定性。我们通过巨灾风险和风险对冲建模示例来阐明该方法。

关键词：因果解释，气候风险，生成式人工智能，符号回归，复杂性指标

Introduction

Autonomous scientific discovery relies on explainable model generation. To arrive at a scientific consensus and describe systems at a proper scale, models often become ensembles of models rather than single theories. Following a multi-target modeling approach, we can define general rules of thumb, such as that simple and accurate explanations are better; the same applies to our generated models. The technique that we discuss in this paper produces multi-target ensembles expressed as human-readable mathematical formulas.

Symbolic regression via genetic programming is a generative AI technique ideal for producing empirical, explanatory scientific advancement. We will begin by introducing general concepts about causal explanation from the philosophy of science and notions related to uncertainty and explanation from complex systems engineering. We will then proceed to show the practicalities of the modeling approach using Data Modeler, the evolutionary computing software written in the Wolfram Language. Finally, we will illustrate the method with an example that relates financial returns in an insurance product to time series of ocean surface temperature anomalies.

Scientific explanation

Wesley Salmon, one of the most influential philosophers of scientific explanation, proposed a definition of explanation according to which to explain is to demonstrate the causal processes behind the occurrence of events (Galavotti, 2022). He advocated that human knowledge is uncertain and that causation should be defined probabilistically (Salmon, 2006). His explanation followed Reichenbach's *Principle of the common cause*, which states that "if an

improbable coincidence has occurred, there must exist a common cause."

Models are instrumental in the practice of science. According to Bokulich, model-based explanations are explanations in which the explanans use particular attributes or behaviors seen in an idealized model or computer simulation to explain why the (usually real-world) explanandum phenomena show the characteristics it does (Bokulich, 2017). Those who have defended the explanatory capacity of models have often claimed that additional requirements must be satisfied before a model's presentation of a notable pattern or phenomenon can be considered an accurate explanation of its real-world counterpart. Not all models are explanatory, and a sufficient theory of model explanation must offer justification for such distinctions. The method of Bokulich is based on Woodward's (Woodward, 2004) counterfactual theory of explanation, in which "the explanation must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways." She contends that model explanations often have three characteristics: The explanans begins with a crucial reference to a scientific model, which, like all models, is an idealized, abstracted, or dramatized portrayal of the target system. Second, the model explains the explanandum by demonstrating how the model components accurately represent the patterns of counterfactual reliance in the target system, allowing one to answer a wide variety of what Woodward refers to as "what-if-things-had-been-different" inquiries. Lastly, there must be what Bokulich refers to as a justificatory phase, which specifies the sphere of application of the model and demonstrates where and to what extent the model can be relied upon as a sufficient representation of the target for the intended purpose(s). Therefore, a necessary condition for modeling is trustability.

As Reichenbach stated, models act as representations and should explain the common causes of things. However, as Tversky puts it, "Typically, there is no single right representation exactly because different representations capture different information, highlight different relationships, and encourage different inferences" (Magnani & Bertolotti, 2017) . In computer modeling and modeling of complex systems, in particular, the answer to the issue of whether computer simulations and analytical models represent distinct modes of thinking depends on the degree of analysis considered. The distinctions between analytical and simulated models become increasingly apparent upon closer inspection. Complex systems are well suited for computer simulations, despite the fact that this may decrease the need for analytical answers (Basso et al., 2017).

Complexity as a measurement of uncertainty

Yaneer Bar-Yam, the complex systems theorist, summarizes the explanation issue in an interesting way: "The inherent compression in the use of language for describing

familiar complex systems is the greatest contributor to uncertainty in complexity estimates" (Bar-yam, 2019). As he goes to show, when characterizing a system n (x,t), we are interested in macroscopic observations throughout time. As with positional uncertainty, a macroscopic observer is unable to discern the time of observation to an accuracy of less than T = [Delta]t. To explain this, we say the system is represented by an ensemble with probability PL, T (n(x;t)) or, more broadly, PL, T (n(x, p;t)). This ensemble consists of all microstates that occur within the time period T. This may look distinct from the spatial uncertainty definition we used.

Nevertheless, the definitions may be rewritten to make them look comparable. In this restatement, we acknowledge that the observer conducts measurements that are, in essence, averages of the different potential microscopic measurements. The observer must characterize the system (or system ensemble) based on the average data across space and time. The usage of an ensemble is advantageous since one observer may only measure one quantity, but other quantities that can be measured with the same degree of accuracy may be considered. For instance, the observer may quantify correlations between the locations of particles that remain constant across time. If the density n(x,t) was averaged across time, these correlations might vanish due to the movement of the whole system. However, when the ensemble is averaged, they do not. The complexity profile C (L, T) is defined as the quantity of information required to describe the ensemble PL, T (n(x;t)).

Evolutionary Model Complexity and Scientific Explanation

So far, we know that there are inherent limitations imposed by language (i.e., by the length of expressions of the mathematical language used to construct explanatory models). We will now discuss a class of generative AI models based on evolutionary algorithms, which offer a viable alternative to create expressions of different degrees of complexity and explanatory power. The implementation is done in the software Datamodeler (Kotanchek, 2010) and has been used to generate scientific models in different disciplines, from life (Pradhan et al., 2020) to social sciences (Venegas et al., 2022).

Definitions

Symbolic regression

A function that executes symbolic regression(s) via genetic programming to identify functional forms, $f[x] = y$ where "x" is an input data record and "y" is a corresponding scalar element of the response vector against which the evolving functions will be judged. The input data may either be a matrix or a vector, but the length must match that of the supplied response vector.

Modeling Objective

A modeling option associated with SymbolicRegression defines the quality of the model, whereas a better model

should have lower metric values. This should be a pure function or a list of pure functions where the inputs are defined as pureFunc[model, modelResponseVector, observedResponseVector, opts].

Model Complexity

A function that returns the model complexity. For a Genetic Programming Model, it is defined as the Total of the LeafCounts of all nodes in the genome. This metric could also be viewed as the visitation length totaling the number of links traversed starting from the root node to each of the end nodes. It is used as part of the default ModelingObjective for symbolic regression.

During SymbolicRegression we typically want to minimize both prediction error and the complexity of the developed models (as well as, possibly, the nonlinearity and other model characteristics). The ModelComplexity is correlated to but not directly coupled to - the nonlinearity of the model. However, it is relatively efficient to calculate. ModelComplexity provides a finer-grained assessment of the complexity of the model structure than the ModelGenome Depth or LeafCount.

The model complexity may be viewed as the visitation length, which is defined as the number of nodes transited from the root node to each of the nodes (not just leaves). The RandomModels below (Figure 1) are labeled with their associated ModelComplexity. To illustrate, a model with two nodes has a ModelComplexity of three—one for the root node plus two for the path from the root node to the other node.

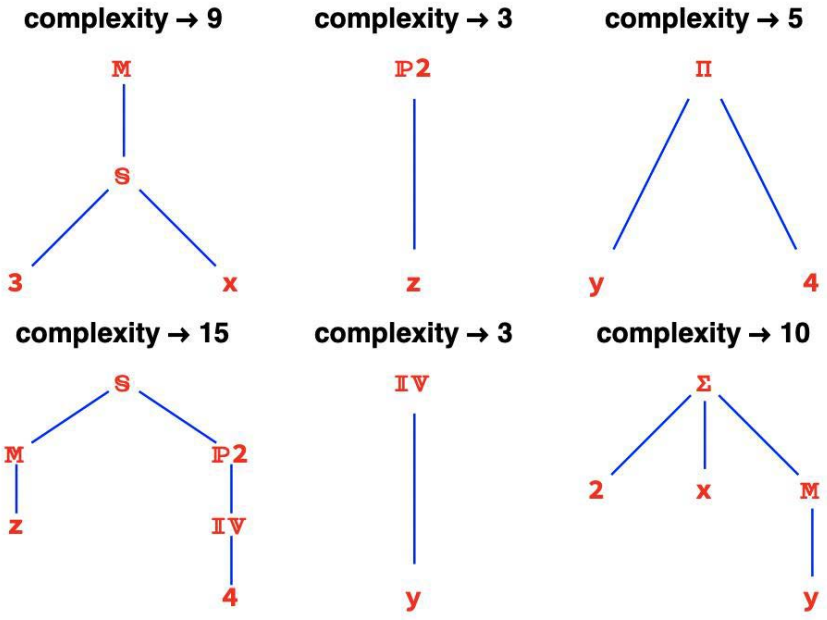


Figure 1. Model tree plot

Genome Complexity

Genome Complexity is a function that returns the genome complexity of the supplied expression, which is defined as the Total of the leaf counts of all nodes in the genome expression. Another way of expressing this is as the visitation length — i.e., the total number of segments that would be traversed starting from the root node to all of the leaves.

Below we synthesize some RandomModels and look at the expressions (phenotypes) as well as the genetic code (genome) used to create that expression. We also visualize the genetic code and calculate its complexity. The ModelComplexity function simply extracts the genetic code from each model and returns its GenomeComplexity (Figure 2).

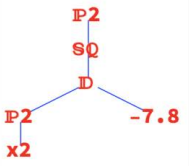
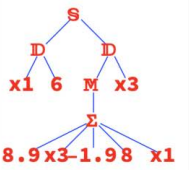
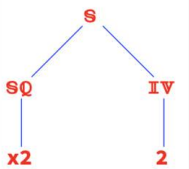
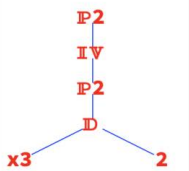
ModelExpression	ModelTreePlot	ModelComplexity	ModelGenome
$-0.13\,x_2^2$		19	ℙ2[𝕊ℚ[𝔻[ℙ2[x2], −7.79098]]]
$\frac{x_1}{6} - \frac{-15.03 - x_1 - x_3}{x_3}$		46	𝕊[𝔻[x1, 6], 𝔻[𝕄[Σ[8.93651, x3, −1.9087, 8, x1]], x3]]
$-\frac{1}{2} + \sqrt{x_2}$		11	𝕊[𝕊ℚ[x2], 𝕀𝕍[2]]
$\frac{16}{x_3^4}$		20	ℙ2[𝕀𝕍[ℙ2[𝔻[x3, 2]]]]

Figure 2. Model complexity

There are several important points to note in the above figure:

(1) The ModelComplexity is calculated based upon the genetic code and NOT the resulting expression. Due to introns (nonfunctional genetic material), we could have the same ModelExpression resulting from models having different genetic material and different GenomeComplexity.

(2) The complexity measure is a measure of representational complexity and NOT the complexity or nonlinearity of the response surface. Thus we could have two expressions of similar complexity that are very different in terms of their nonlinearity, e.g., IV[x] → 1/x

and SQ[x] → SQRT(x) which would both have a genetic complexity of 3 but be much different in terms of their response characteristics—especially if in a region including zero.

ModelEnsemble

ModelEnsemble is a data structure that represents a model comprised of a group of models. Typically created using CreateModelEnsemble or CreateFittedEnsemble , the ensemble is evaluated based upon the EnsembleEvaluationFunction with the trustability of that prediction measured using the EnsembleDivergenceFunction.

To start, let us define a reference function and sample it at a variety of points, as seen in Figure 3.

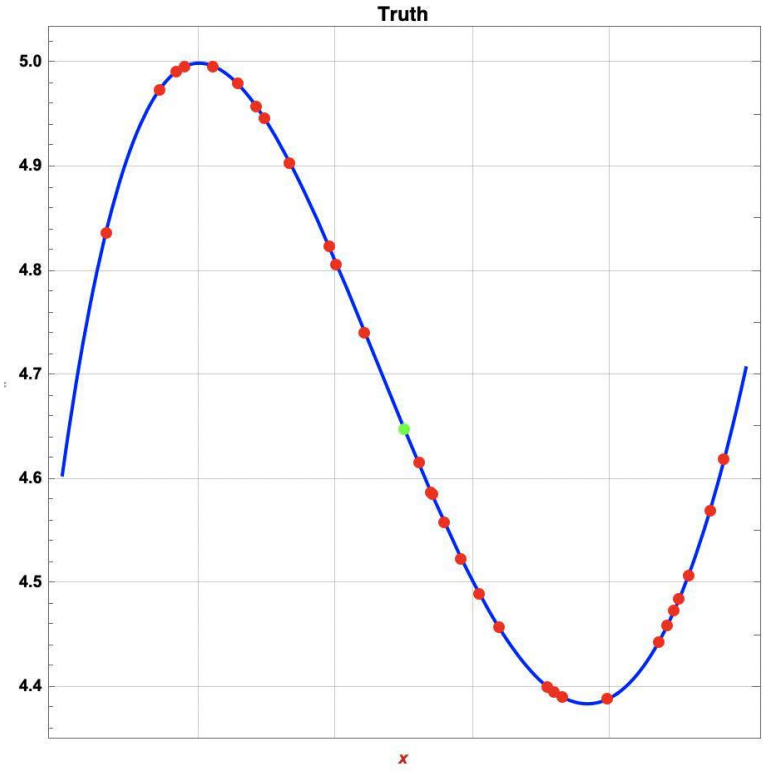


Figure 3. "Truth" model

To build an ensemble, we want to have a diverse collection of models. Towards that end, we will run eight IndependentEvolutions in parallel. One thousand thirty-four models are generated, as shown in Figure 4.

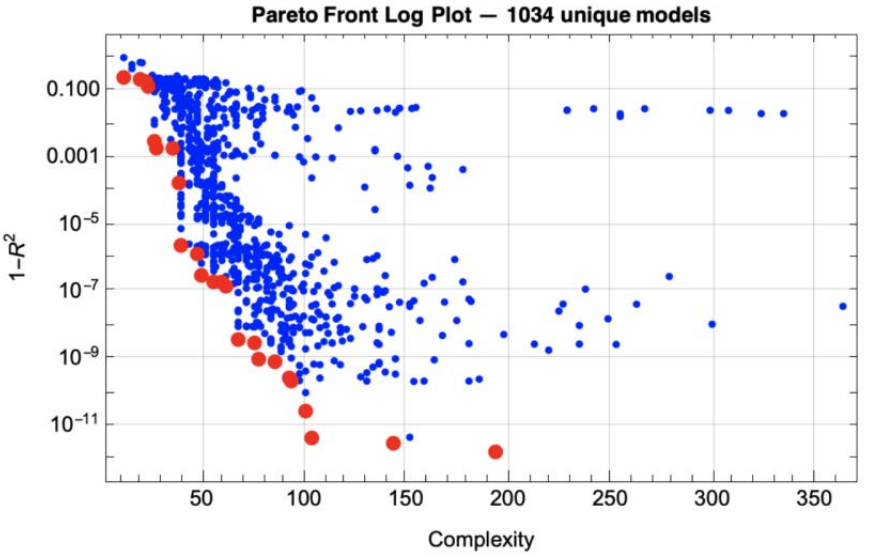


Figure 4. Accuracy/Complexity tradeoff for models

Next, we use CreateModelEnsemble to select an ensemble from the identified region of good models (Figure 5). The default behavior of this function is to overweight the knee of the ParetoFront while simultaneously including diverse models.

Model Selection Report

	Complexity	$1-R^2$	Function
1	26	0.003	$4.99+0.43\,x_1^2\,(-2.13+x_1)$
2	47	3.180×10^{-5}	$140.01-\frac{7.43\times10^{-5}}{x_1}-23.05\,x_1+2.86\,x_1^2-\frac{790.43}{5.85+x_1}$
3	61	1.488×10^{-7}	$13.41-3.42\,x_1+0.29\,x_1^2+0.20\,x_1^3-\frac{41.31}{4.91+2\,x_1}$
4	67	3.799×10^{-9}	$79.69-13.19\,x_1+1.46\,x_1^2+0.12\,x_1^3-\frac{1.11}{1.40+x_1}-\frac{432.62}{5.85+x_1}$
5	67	1.256×10^{-8}	$8.86-1.93\,x_1-0.14\,x_1^2+0.29\,x_1^3-(9.91\times10^{-3})\,x_1^4-\frac{7.71}{2+x_1}$
6	67	2.885×10^{-8}	$44851.59-962.48\,x_1+20.48\,x_1^2-0.16\,x_1^3-\frac{9.06}{2.07+x_1}-\frac{2093801.60}{46.69+x_1}$
7	67	3.594×10^{-7}	$23.76-6.25\,x_1+0.98\,x_1^2+(7.22\times10^{-2})\,x_1^3+(1.34\times10^{-2})\,x_1^4-\frac{56.29}{3+x_1}$
8	75	3.023×10^{-9}	$79.50+\frac{1.23\times10^{-6}}{x_1}-13.16\,x_1+1.46\,x_1^2+0.12\,x_1^3-\frac{1.12}{1.40+x_1}-\frac{431.52}{5.85+x_1}$
9	76	3.554×10^{-9}	$23007.56-494.92\,x_1+10.70\,x_1^2-\frac{10.17}{2.07+x_1}-\frac{1073867.00}{46.69+x_1}+\frac{8.95}{9.10+x_1^2}$
10	77	9.883×10^{-10}	$81.10-13.21\,x_1+1.44\,x_1^2+0.12\,x_1^3-\frac{449.51}{5.98+x_1}-\frac{5.54}{5.85+4\,x_1}$
11	77	1.169×10^{-8}	$8.81-1.91\,x_1-0.14\,x_1^2+0.29\,x_1^3-(1.01\times10^{-2})\,x_1^4-\frac{22.79}{5.98+3\,x_1}$
12	143	3.040×10^{-12}	$-926.48-50.96\,x_1-2.84\,x_1^2+(2.22\times10^{-2})\,(8.04-x_1)\,x_1^3-\frac{17649.70}{-18.90+x_1}-\frac{0.79}{(3+2\,x_1)^4}-\frac{148.86}{(8.33+3\,x_1)^2}$

Figure 5. Mathematical expressions of models

As we can see from the EnsemblePredictionPlot (Figure 6), the ensemble prediction fits the data quite well.

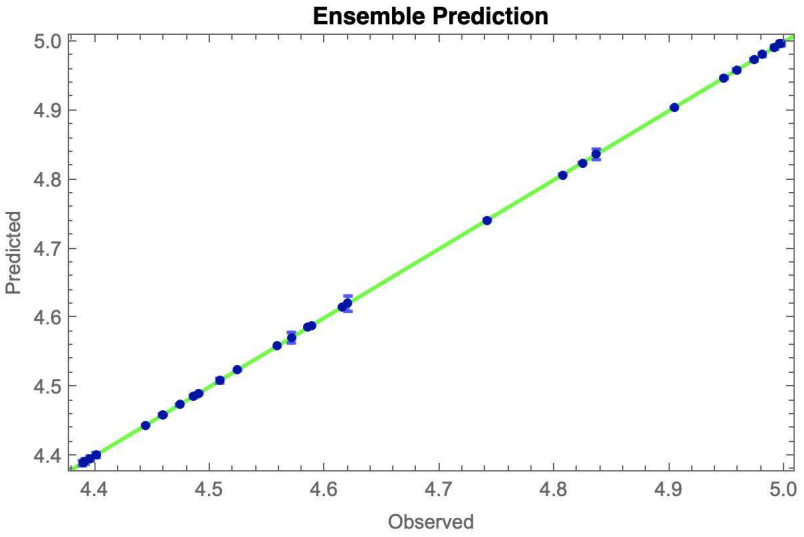


Figure 6. Prediction plot for models in ensemble

The EnsembleQuality is calculated and embedded in the generated ModelEnsemble. Here it is displaying {R2, # Variables ,# Models, Avg Model Complexity} = {1., 1, 12, 70.8333}

Next, let us look at the ensemble prediction relative to the available data and the (normally unknown) true model. Again, we see that the ensemble prediction (Figure 7) fits the observed data well and, additionally, has reasonable behavior when asked to extrapolate.

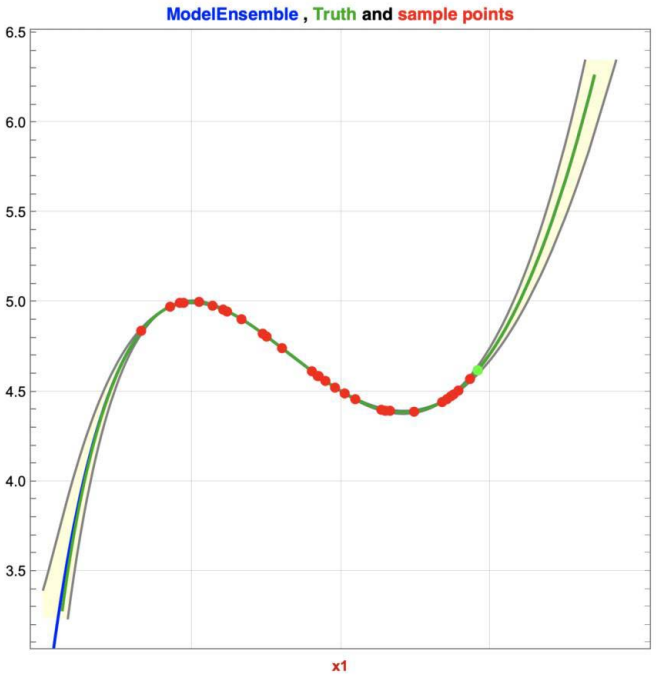


Figure 7. Comparison using average of models in the ensemble

If we include the behavior of the individual models embedded in the ensemble, we see that they are constrained to agree where there is data; however, outside that region, they are encouraged to deviate due to the emphasis on diversity. Figure 8 shows both the average and the individual models.

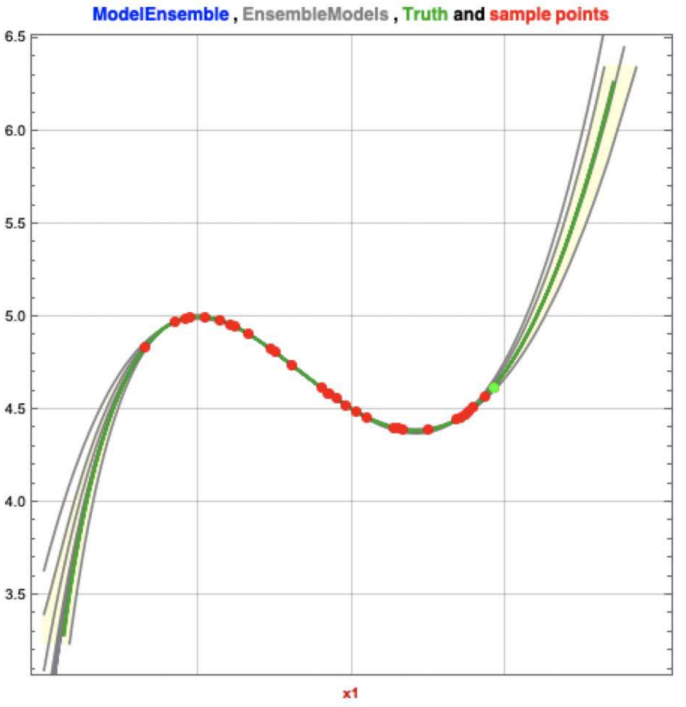


Figure 8. Comparison using constituent models of the ensemble

Model diversity is the foundation for detecting the trustability of the ensemble prediction. As we can see below, where there is data, the models agree, and—where not constrained by the data—the constituent models of the ensemble diverge.

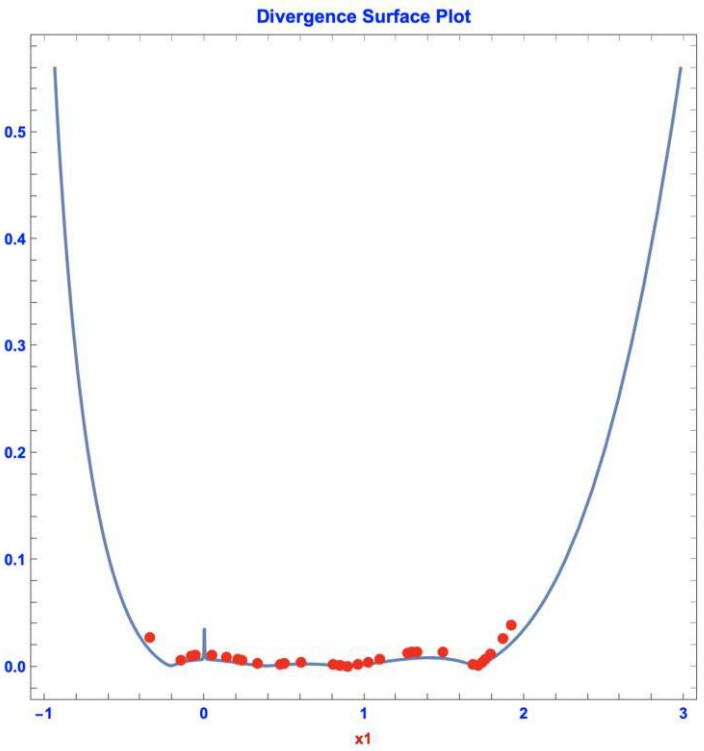


Figure 9. Divergence

Uncertainty, Chaotic Systems, and Complexity

The divergence in an ensemble of evolutionary models is the key to assessing trustability in scientific modeling. Arguably, it is even more useful when dealing with complex systems modeling, as we will show.

According to Palis (ROGERS, 1999), sensitivity with regard to initial conditions is crucial: the long-term outcome may vary significantly when just little changes are made to the original conditions or first event. This is evidently the case with mathematical models for weather forecasting, as Lorenz pointed out in his astounding 1963 work: Future responses change significantly (degree of uncertainty) with very, very tiny variations in beginning data. Such variance is inherent since it is difficult to predict the precise beginning values of temperature, pressure, the quantity of precipitation, and so forth. These systems are termed chaotic.

The notion of a complex system is significantly more modern. We want to describe extremely complex systems, such as the behavior of neuron networks (brain), and we know some of the properties that we want to impose on the dynamical systems that could model them, such as nonlinearity, adaptability (i.e., the system is constantly changing in response to external parameters),

randomness, multiple attractors, fractal structure, and others. Also, such a system should not be completely chaotic, but it should be close (on the borderline of chaotic systems). Particularly, we should not have exponential sensitivity with regard to initial conditions for a complex system. In other words, the long-term behavior of dynamical systems that are suitable for modeling a "complex phenomena" should be sensitive to initial conditions but not as strongly as in chaotic systems. In addition, the local (almost punctual) structure of the dynamical system should be simple and resilient, exhibiting little variation when the system is modified significantly.

As Palis puts it, mathematically, we are far more advanced in our understanding of chaotic systems (there is still a great deal of work to be done, but we can at least propose a possible global scenario) than complex systems, primarily due to the lack of good dynamical models for the examples, such as the brain, that we have considered as complex phenomena. Here is where generative models (of the evolutionary type) could really make a difference in our understanding of complex systems.

A brief empirical model example

We will now analyze a complex hybrid system (economic) and a chaotic system (weather). Specifically, we use investment data (hedge fund) generated by the interaction of economic agents (investors) placing bets on insurance models of the weather. To illustrate the importance of scientific modeling related to this topic, it should suffice to consider that the particular fund manager, in this case, currently has 34 billion USD under management; this class of modeling will become increasingly important due to climate change risk and the need to better understand and mitigate its effects.

The sources are as follows:

Weather data: is NOAA National Centers for Environmental Information, Climate at a Glance: Global Time Series, published April 2022, retrieved on May 1, 2022, from https://www.ncdc.noaa.gov/cag/

Insurance hedge-fund data: Returns from Eurekahedge for the following product *PIMCO ILS Fund SP II* and prospectus https://www.pimco.com/en-us/insights/investment-strategies/featured-solutions/insurance-linked-securities-seeking-returns-beyond-traditional-assets/

Description

We generated models using as predictor variables ocean temperatures (globally and for 14 regions of the world). The target variable is the monthly returns of the insurance-linked hedge fund for the period from December 2019 to March 2022 (27 observations). In total, 1069 models were generated.

Model Selection Report (Returns)

	Complexity	$1-R^2$	Function
1	22	0.659	$0.28 - \frac{\text{EastPac } 0.28}{\text{Caribbean}^4}$
2	29	0.583	$0.34 - (-\text{Caribbean} + \text{Global})^4\ 153.69$
3	32	0.578	$1.03 - \text{Atlantic}^2\ (1 - \text{Caribbean})^2\ 45.16$
4	36	0.491	$0.56 - \text{Atlantic}^2\ (-\text{Caribbean} + 0.96)^4\ 339.64$
5	39	0.447	$0.59 - \frac{\text{Atlantic } (-\text{Caribbean}+0.96)^4\ 301.90}{\text{Asia}}$
6	54	0.441	$0.57 - \text{Africa Atlantic}\,(1 - \text{Caribbean})^2 \left(\sqrt{\text{Atlantic}} - \text{Caribbean}\right)^2\ 601.91$
7	55	0.441	$0.68 - \text{Atlantic}^2 \left(\text{Atlantic}^{1/4} - \text{Caribbean}\right)^4\ 1075.76$
8	57	0.431	$0.65 - \text{Atlantic}\left(\sqrt{\text{Atlantic}} - \text{Caribbean}\right)^2 \left(\sqrt{\text{Caribbean}} - \text{Caribbean}\right)^2\ 4261.01$
9	60	0.415	$0.66 - \text{Africa Atlantic}\left(\sqrt{\text{Atlantic}} - \text{Caribbean}\right)^2 \left(\sqrt{\text{Caribbean}} - \text{Caribbean}\right)^2\ 3306.64$
10	65	0.411	$0.63 - \frac{\text{Atlantic}^2\ (-\text{Caribbean}+0.96)^4\ 1314.45}{2\ \text{Asia}+\text{Caribbean}}$
11	66	0.387	$0.56 - \frac{\text{Atlantic } (-\text{Caribbean}+0.96)^4\ 1744.53}{\text{Asia}^4+\text{Europe}+0.96}$
12	72	0.386	$0.58 - \frac{\text{Atlantic } (-\text{Caribbean}+0.96)^4\ 1560.15}{\text{Asia}^4+\text{EastNPac}^2+\text{Europe}}$
13	86	0.385	$0.58 - \frac{\text{Atlantic } (-\text{Caribbean}+0.96)^4\ 1512.53}{\text{Asia}^4+\text{Europe}+(-\text{Caribbean}+0.96)^2}$
14	102	0.362	$0.66 - \frac{\text{Atlantic}^2\ (1-\text{Caribbean})^4\ 2874.69}{5+\left(\text{Asia}+\text{Caribbean}-\frac{1}{\text{Hawaii}}\right)^2+\text{Oceania}}$
15	110	0.349	$0.62 - \frac{\text{Atlantic}^2\ (1-\text{Caribbean})^4\ 2496.48}{5+\left(\text{Asia}^2+\text{Caribbean}-\frac{1}{\text{Hawaii}}\right)^2+\text{Oceania}}$
16	116	0.345	$0.60 - \frac{\text{Atlantic}^2\ (1-\text{Caribbean})^4\ 2251.07}{-\text{Caribbean}+\left(\text{Asia}^2+\text{Caribbean}-\frac{1}{\text{Hawaii}}\right)^2+5.96}$
17	178	0.337	$0.53 - \frac{\text{Atlantic}^2\ 753990.13}{\left(-\text{Atlantic}+\frac{1}{\text{Atlantic}^2\ (1-\text{Caribbean})^4}\right)^2 \left(5+\text{Caribbean}+\left(\text{Asia}^2+\text{Caribbean}-\frac{1}{\text{Europe}}\right)^2\right)}$
18	185	0.329	$0.56 - \frac{\text{Atlantic}^2\ 726417.62}{\left(5+\text{Caribbean}+\left(\text{Asia}^2+\text{Caribbean}-\frac{1}{\text{Europe}}\right)^2\right) \text{Global} \left(\frac{1}{\text{Atlantic}^2\ (1-\text{Caribbean})^4}+2.19\right)^2}$
19	247	0.322	$0.56 - \frac{\text{Atlantic}^2\ 9950852.90}{\left(-1+\frac{1}{\text{Atlantic}^2\ (1-\text{Caribbean})^4}\right)^2 \left(5+\text{Caribbean}+\left(\text{Asia}^2+\text{Caribbean}-\frac{1}{\text{Hawaii}}\right)^2\right) \left(\text{Atlantic}-\left(\frac{1}{\text{Global}}-6.75\times10^{-2}\right)7.23+11.33\right)^2}$

Figure 10. Subset of models (first round of modeling), complexity and error

Discussion

Figure 10 shows models ordered from less complex (rank 1) to higher complexity (rank 19). Models with similar mathematical expressions, in terms of descriptive form, have close complexity metrics (for instance, for rank 10, complexity is 65; for rank 11 is 66). Although it may appear that the error of the models in the sample is relatively high (between 0.322 and 0.659), this is not an issue because the purpose of the modeling in a preliminary exploratory round is to assess modeling potential and explanatory causality, not generating predictive approximations.

The first modeling attempt is quick (only 5 minutes); however, it renders insights almost immediately. We note how one region of the world is no-

toriously absent from the list of models: North America. In fact, if we investigate the presence of variables across models (Fig 11), we find that Caribbean ocean temperatures are a factor in 100% of the models while North America only has about 1%.

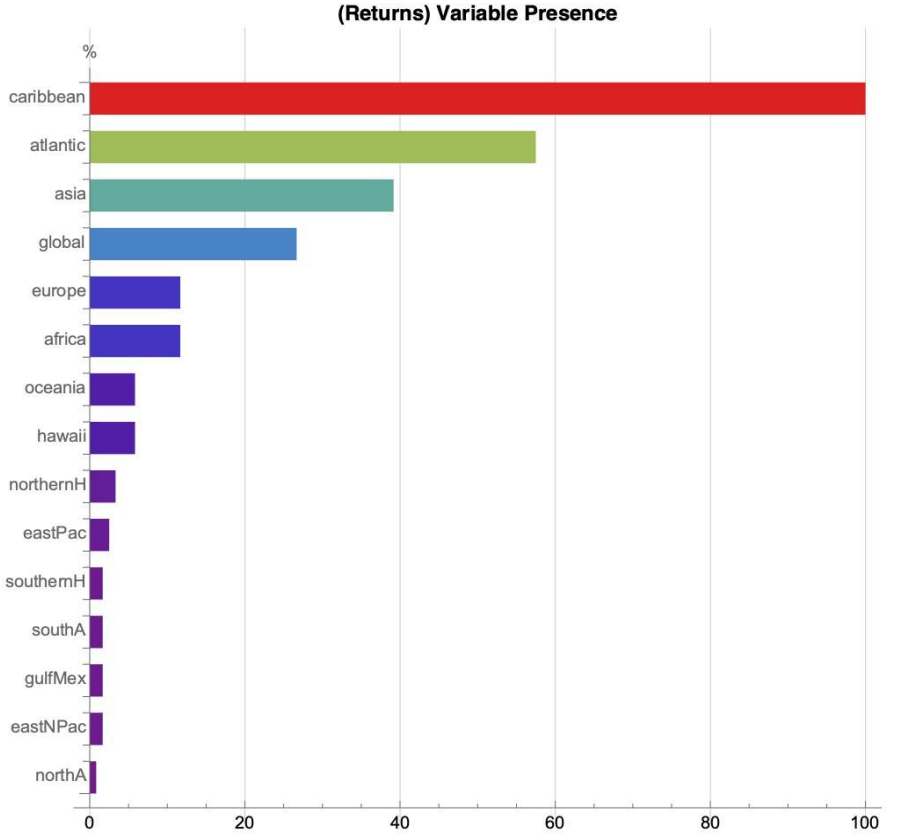


Figure 11. Variable presence (percentage) across models

At this point, it is insightful to consider the process of the fund operator. As indicated in the prospectus, they use multiple catastrophe-risk models augmented by proprietary analytics to better comprehend, quantify, and manage risks based on meteorological science (floods, hurricanes, etc.), property engineering, and claims data.

Arguably, just one round of modeling is necessary (as a toy model) but not sufficient to characterize the system. In the Appendix, we include exhibits that describe the results of a second round of modeling, although many more are encouraged in real-world applications. Another reason why multiple rounds of modeling are advisable is that we should not draw quick conclusions from a stochastic modeling process; for instance, the prevalence of the Caribbean region and exclusion of North America in the first round of modeling might as well be a by-product of the way particular modeling settings handle exclusions.

Uncertainty reduction by Ensemble creation

If one wants to increase the richness of the explanation, it suffices to create a group of models (en ensemble) using the appropriate combination of good performance (low error) and low complexity. Such a group is marked by the square box in Figure 12 and then shown in more detail in Figure 13.

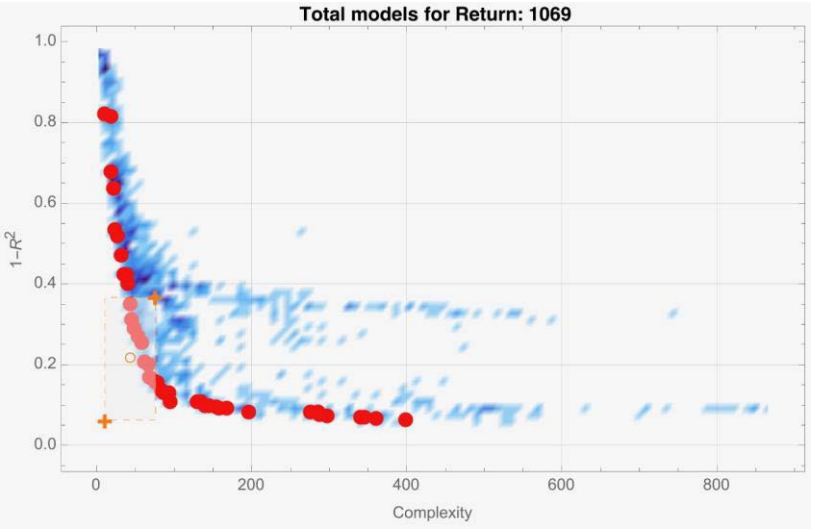


Figure 12. Optimal models

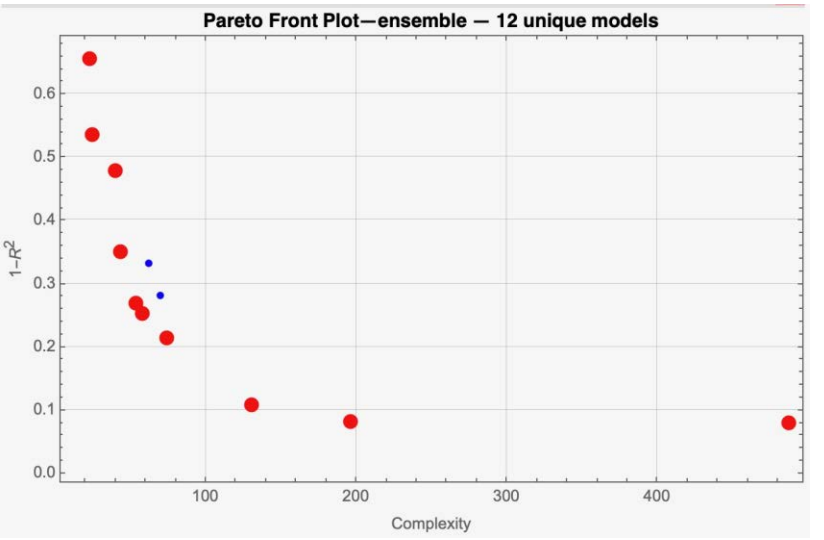


Figure 13. Ensemble sub sample

The qualities of the ensemble are superior to any single model's. Figure 14 shows the key statistics and variables for our example, an ensemble containing 12 models.

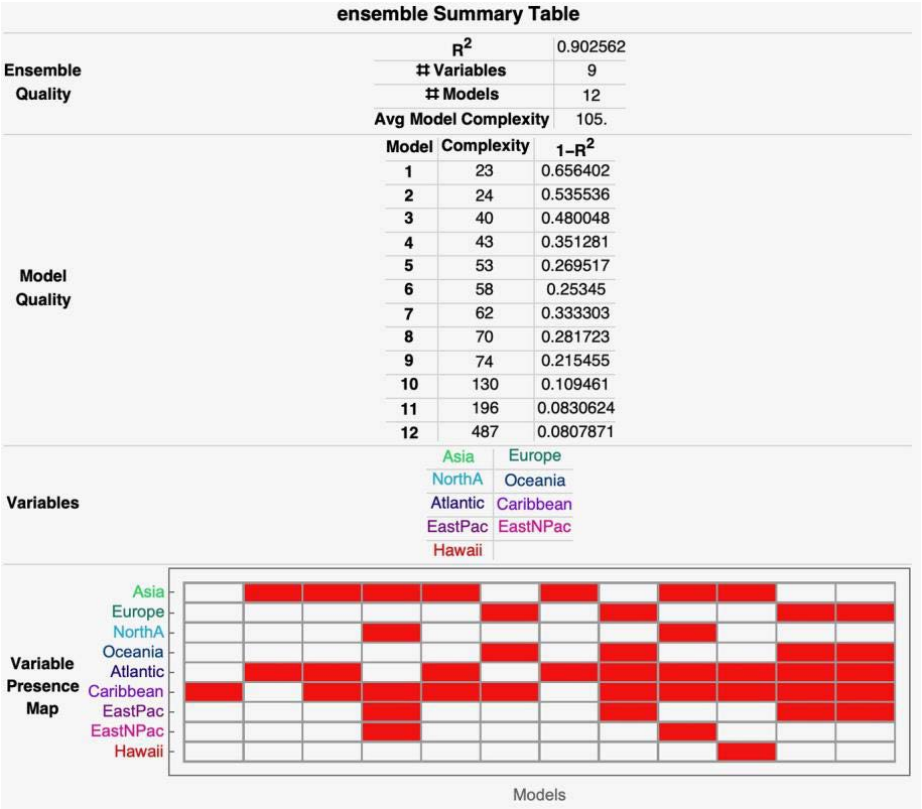


Figure 14 .Ensemble summary

As we can see in the tables, the quality (R2=0.92562) of the ensemble is higher than the accuracy of any individual model. At the same time, the different complexity values provide access to different levels of explanation. Therefore, we gain diversity while retaining the descriptive benefits of the average in the ensemble.

Arguably, just one round of modeling is necessary (as a toy model) but not sufficient to characterize the system. In the Appendix, we include exhibits that describe the results of a second round of modeling, although many more are encouraged in real-world applications. Another reason why multiple rounds of modeling are advisable is that we should not draw quick conclusions from a stochastic modeling process; for instance, the prevalence of the Caribbean region and exclusion of North America in the first round of modeling might as well be a by-product of the way particular modeling settings handle exclusions (i.e., of some of the nuances of small data sets as well as ensemble definition).

Conclusion

We have shown how a class of generative AI models based on evolutionary algorithms allows us to be aware of and ultimately reduce uncertainty. We become more

confident of the causal explanations (driver variables) of particular phenomena, and we can ascribe both a complexity metric and a quality measurement to each model (or group of models) developed. We can also include chaotic and complex systems in the analysis, using summary figures and statistics that describe the behavior and interrelation of physical and social systems.

The policy implications are evident: instead of long policy development cycles that depend on years-long research programs, the practice of scientific discovery becomes a highly iterative process where relevant questions and the possible causes of things surface quickly.

Appendix

A second round of modeling might include a subset of popular variables. For instance, the combination *of Asia, the Atlantic, the Caribbean, and Hawaii* is present in 6.8% of the models of the first round (Fig 15).

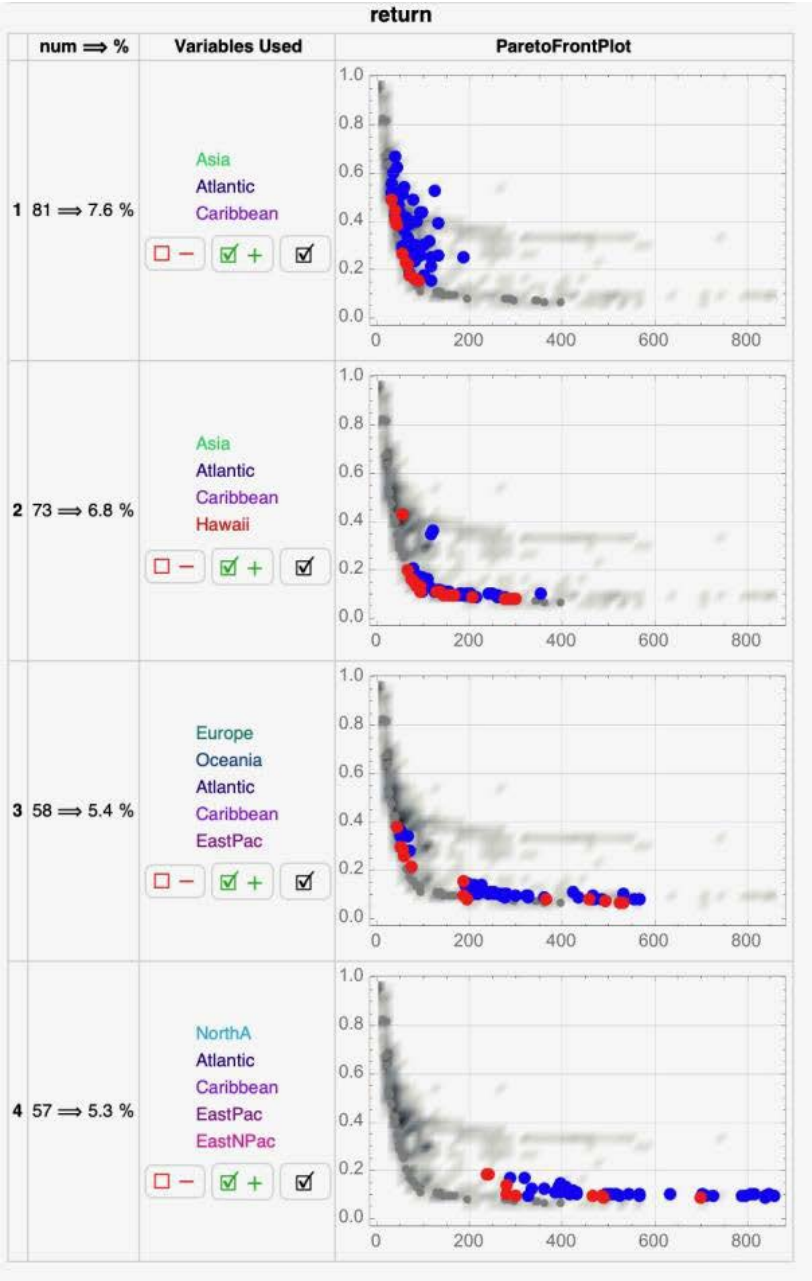


Figure 15. Most popular variable combinations

The second round produced 2076 unique models. The settings are shown below (Fig 16).

Return

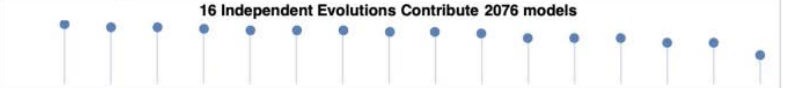
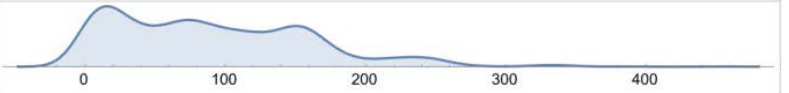
Control	Count ⟹ Settings
Searches	**16 Independent Evolutions Contribute 2076 models** 
Model Age	
TimeConstraint	2076 ⟹ 300
FunctionPatterns	2076 ⟹ {Plus, Times, Divide, Subtract, xx^2, $\sqrt{xx}$, $\frac{1}{xx}$, Minus}
MetaVariables	2076 ⟹ None
CategoryThreshold	2076 ⟹ {3, 0.9}
ConvertedForms	2076 ⟹ All
BinaryFormat	2076 ⟹ ZeroSymmetric
ResponseMapping	2076 ⟹ {}
InputMapping	2076 ⟹ {{1, TimeData, {time$_DateObject ↣ QuantityMagnitude[DateDifference[DateObject[{2020, 1, 1, 0, 0, 0.}, Instant, Gregorian, −6.], time$, Month]]}}}
OptimizeLinearModel	1290 ⟹ True 786 ⟹ False
RobustModels	2076 ⟹ False
BasisSetLimit	1803 ⟹ 5 273 ⟹ 3
RescaleData	2076 ⟹ None
NumberOfVariables	2076 ⟹ ∞
PowerLimit	2076 ⟹ 4
AllowedVariables	2076 ⟹ {asia, atlantic, caribbean, hawaii}
ExcludedVariables	2076 ⟹ Complement
ModelingObjective	2076 ⟹ {ModelComplexity, ModelAccuracy}
ModelingObjectiveNames	2076 ⟹ {Complexity, $1-R^2$}
CreateBasisSetModel	1290 ⟹ True 786 ⟹ False
EvolutionStrategy	2076 ⟹ Automatic
DataSubsetSize	2076 ⟹ All
DataSubsetSelectionFunction	2076 ⟹ RandomSample
DataSegments	2076 ⟹ Automatic
ActiveGenomeSimplification	2076 ⟹ False
AlignModel	2076 ⟹ OptimizeLinearModel
NumericInputRequirement	2076 ⟹ 0.75
NumericPredictionRequirement	2076 ⟹ 1.
ProjectName	2076 ⟹ insurancehedge
FileName	2076 ⟹ 2−return
RoundName	2076 ⟹ Round02
TargetColumn	2076 ⟹ 2
TargetLabel	2076 ⟹ Return

Figure 16. Model summary

As usual, ensembles are created in subsequent rounds of modeling. Typically, both the overall modeling performance and individual variable behavior are analyzed, as seen in Fig 17.

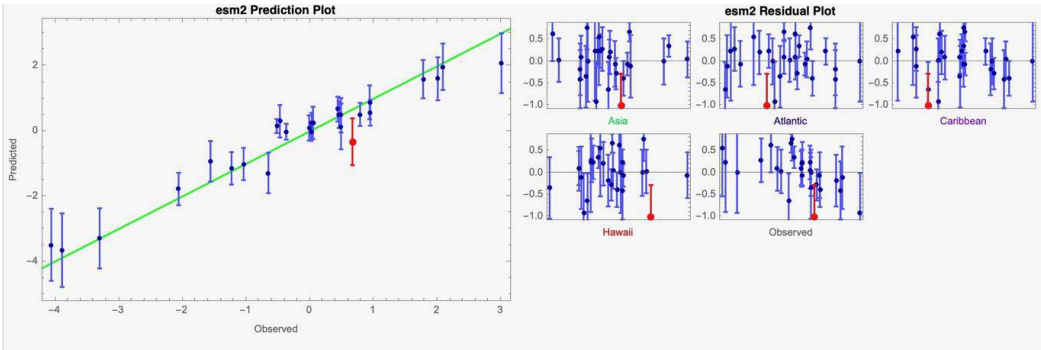


Figure 17. Ensemble prediction plot and residuals

References

Salmon, W. C. Four Decades of Scientific Explanation. (2006). University of Pittsburgh Press. https://doi.org/10.2307/j.ctt5vkdm7

Bokulich, Alisa. Models and Explanation. (2017). In *Springer Handbook of Model-Based Science* (pp. 103–118). Springer International Publishing. https://doi.org/10.1007/978-3-319-30526-4_4

Woodward, James.The Causal Mechanical and Unificationist Models of Explanation. (2004). In *Making Things Happen* (pp. 350–373). Oxford University Press New York. https://doi.org/10.1093/0195155270.003.0008

Springer Handbook of Model-Based Science. (2017). Springer International Publishing. https://doi.org/10.1007/978-3-319-30526-4

Basso, Alessandra; Lisciandra, Chiara; Marchionni, Caterinav. Hypothetical Models in Social Science. (2017). In *Springer Handbook of Model-Based Science* (pp. 413–433). Springer International Publishing. https://doi.org/10.1007/978-3-319-30526-4_19

Bar-yam, Yaneer. Dynamics Of Complex Systems. (2019). CRC Press. https://doi.org/10.1201/9780429034961

Kotanchek, M. Real-world data modeling. (2010). *Proceedings of the 12th Annual Conference Comp on Genetic and Evolutionary Computation - GECCO 10*. https://doi.org/10.1145/1830761.1830921

Pradhan, Pallab; Chatterjee, Paramita; Stevens, Hazel Y.; Glen, Chad; Medrano-Trochez, Camila; Jimenez, Angela; Kippner, Linda; Seeto, Wen-Jun; Li, Ye; Gibson, Greg; Kurtzberg, Joanne; Kontanchek, Theresa; Yeago, Carolyn; Roy, Krishnendu. *Single-Cell Transcriptomic Attributes and Unbiased Computational Modeling for the Prediction of Immunomodulatory Potency of Mesenchymal Stromal Cells*. (2020). https://doi.org/10.1101/2020.09.12.294850

Venegas, Percy; Britez, Isabel; Gobet, Fernand. Ensemble Models Using Symbolic Regression and Genetic Programming for Uncertainty Estimation in ESG and Alternative Investments. (2022). In *Big Data in Finance* (pp. 69–91). Springer International Publishing. https://doi.org/10.1007/978-3-031-12240-8_5

Rogers, C. L.: World Conference on Science for the Twenty-First CenturyA New Commitment. (1999). *Science Communication*, *21*(2), 179–182. https://doi.org/10.1177/1075547099021002006