TRACKING CLIMATE MODELS

CLAIRE MONTELEONI*, GAVIN SCHMIDT**, AND SHAILESH SAROHA***

Abstract. Climate models are complex mathematical models designed by meteorologists, geophysicists, and climate scientists to simulate and predict climate. Given temperature predictions from the top 20 climate models worldwide, and over 100 years of historical temperature data, we track the changing sequence of which model currently predicts best. We use an algorithm due to Monteleoni and Jaakkola that models the sequence of observations using a hierarchical learner, based on a set of generalized Hidden Markov Models (HMM), where the identity of the current best climate model is the hidden variable. The transition probabilities between climate models are learned online, simultaneous to tracking the temperature predictions. On historical data, our online learning algorithm's average prediction loss nearly matches that of the best performing climate model in hindsight. Moreover its performance surpasses that of the average model prediction, which was the current state-of-the-art in climate science, the median prediction, and least squares linear regression. We also experimented on climate model predictions through the year 2098. Simulating labels with the predictions of any one climate model, we found significantly improved performance using our online learning algorithm with respect to the other climate models, and techniques.

1. Introduction

The threat of climate change is one of the greatest challenges currently facing society. With the increased threats of global warming, and the increasing severity of storms and natural disasters, improving our understanding of the climate system has become an international priority. This system is characterized by complex and structured phenomena that are imperfectly observed and even more imperfectly simulated. A fundamental tool used in understanding and predicting climate is the use of *climate models*, large-scale mathematical models run as computer simulations. Geophysical experts, including climate scientists and meteorologists, encode their knowledge of a myriad of processes into highly complex mathematical models. One climate model will include the modeling of such processes as sea-ice melting, cloud formation as a function of increased pollution in the atmosphere, and the creation, depletion and transport of many atmospheric gases. These are just a few of the processes modeled in one model; each climate model is a highly complex system.

In recent years, the magnitude of data and climate model output is beginning to dwarf the relatively simplistic tools and ideas that have been developed to analyze them. In this work, we demonstrate the advantage of a machine learning approach, over the state-of-the-art in climate science, for combining the predictions of multiple climate models. In addition to our specific contributions, we encourage the broader study of *climate informatics*, collaborations between climate scientists and machine learning researchers in order to bridge this gap between data and understanding.

The global effort on climate modeling started in the 1970s, and the models have evolved over time, becoming extremely complex. There are currently about 20 laboratories across the world whose climate models inform the Intergovernmental Panel on Climate Change (IPCC), a panel established by the United Nations in 1988, that was recognized for its work on climate change with the 2007 Nobel Peace Prize (shared with former US Vice President Al Gore). Work done to improve the utilization of global climate model predictions would be very significant to the next IPCC report.

*Center for Computational Learning Systems, Columbia University, cmontel@ccls.columbia.edu.
**Center for Climate Systems Research, Columbia University, and NASA Goddard Institute for Space Studies, gschmidt@giss.nasa.gov.
***Department of Computer Science, Columbia University, shaileshsaroha@gmail.com.

Currently there is very high variance among the predictions of these 20 models. This may stem from a variety of reasons. Each was designed from first principles by a different team of scientists, and thus the models differ in many discretization assumptions, as well as in some of the science informing each process modeled. It was observed however, that while the variance is high, the average prediction over all the models is a more consistent predictor (over multiple quantities, such as global mean temperature, performance metrics, and time periods), than any one model [32, 33].

Our contribution is an application of a machine learning algorithm that produces predictions that match or surpass that of the best model for the entire sequence. We use online learning algorithms with the eventual goal of making both real-time and future predictions. Moreover, our experimental evaluations reveal that, given the non-stationary nature of the observations, and the relatively short history of model prediction data, a batch approach has performance disadvantages. Our algorithm achieves lower mean prediction loss than that of several other methods, including prediction with the average over model predictions. This is an impactful result because to date, the average of all models' predictions was believed to be the best single predictor of the whole sequence [32, 33].

Related work in Machine Learning and Data Mining. There are a few other applications of machine learning and data mining to climate science. Data mining has been applied to such problems as mining atmospheric aerosol data sets [31, 30], analyzing the impacts of climate change [20], and calibrating a climate model [6]. Clustering techniques have been developed to model climate data [38]. Machine learning has been applied to predicting the El Niño climate pattern [21], and modeling climate data [39]. In another work, machine learning and data mining researchers proposed the use of data-driven climate models [23]. There has also been work on integrating neural networks into global climate models [19, 18].

We are not aware of applications, beyond our own, of machine learning to the problem of tracking global climate models. Our work builds on our preliminary results which have been workshopped with colleagues in both machine learning and climate science [26, 27, 28]. We apply the Learn-α algorithm of Monteleoni and Jaakkola [25] to track a shifting sequence of temperature values with respect to the predictions of "experts," which we instantiate in this case with climate models. That work extends the literature on algorithms to track a sequence of observations with respect to the predictions of a set of experts, due to Herbster and Warmuth [15], and others.

2. The problem of tracking climate models

2.1. **Climate models.** A fundamental tool used in predicting climate is the use of large-scale physics-based models of the global atmosphere/ocean/cryosphere system. As illustrated in Figure 1, these General Circulation Models (GCMs) simulate the basic processes seen in observations, such as cloud formation, rainfall, wind, ocean currents, radiative transfer through the atmosphere etc., and have emergent properties, such as the sensitivity of climate to increasing greenhouse gases, that are important to making any climate forecasts [36]. It is important to note that unlike the use of the term *model* in machine learning, here we denote systems of mathematical models, that are *not* data-driven. These complex systems are composed of individual mathematical models of each of the processes mentioned, among others. The models are based on scientific first principles from the fields of Meteorology, Oceanography, and Geophysics, among others.

There are a number of challenges in using these models. First, the simulated climate in each model has biases when compared to real world observations. Second, the internal variability seen in these models (more colloquially, the "weather") is not synchronized to the weather in the real world (these models are quite different from the models used for numerical weather prediction on multi-day time scales), and indeed can be shown to have a sensitive dependence to initial conditions (i.e. it is chaotic). Third, each of the models has a different sensitivity to external drivers of climate (such as human-caused increases in greenhouse gases and aerosols, large volcanic eruptions, solar activity

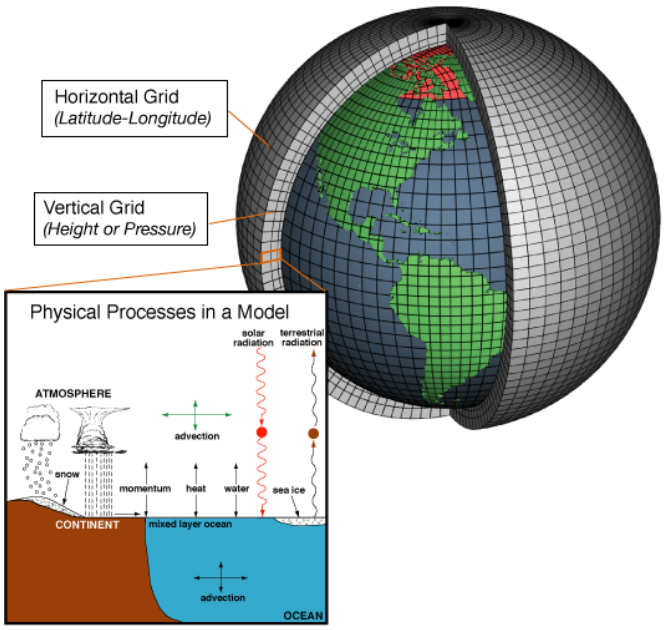

FIGURE 1. Global climate model (schematic due to [1]).

etc.), which is wide enough to significantly affect future projections.[1] Fourth, while robust responses of the modeled climate can be derived from imposing these external drivers of climate, knowledge of those drivers in the past can be uncertain. Thus evaluating the quality of multi-decadal climate projections is fraught with uncertainty.

Any simulation of these models is made up of two elements, the externally forced "climate" signal and the stochastic "internal climate variability." The former can be estimated quite effectively by generating multiple simulations from one individual model, where each simulation has an independent and uncorrelated realization of the internal variability. The real world can be considered as a single realization of its internal variability along with an (uncertain) signal caused by external climate drivers mentioned above. Thus, detection of a climate change and its attribution to any particular cause needs to incorporate the uncertainties in both the expected signal and the internal variability [35].

For projections of future climate, there are three separate components to the uncertainty [14]. First is the scenario uncertainty: the fact that we do not have future knowledge of technological, sociological or economic trends that will control greenhouse gas and other emissions in the future. Given the inertia of the economic system, this uncertainty is small for the next couple of decades, but grows larger through time. The second component of the uncertainty is associated with internal variations of the climate system that are not related to any direct impact of greenhouse gases etc. Such variability is difficult to coordinate between the models and the real world, and the degree to which it is predictable is as yet unclear. This component is large for short time periods but becomes less important as the externally driven signal increases.

[1]In climate science terminology, a climate model *projection* denotes a simulation for the future given a particular scenario for how the external drivers of climate will behave. It differs from a prediction in that a) the scenario might not be realized, and b) only the component of the climate that is caused by these external drivers can be predicted while the internal variability cannot be. Thus projections are not statements about what *will* happen, but about what *might* happen. However we will also use the term *prediction* interchangeably.

The third component, and the one that this paper focuses on, is the uncertainty associated with the models themselves. The relative importance of this is at its maximum between roughly 20 and 50 years into the future (long enough ahead so that the expected signal is stronger than the internal variability, but before the uncertainty in the scenarios becomes dominant). The source of model uncertainties might be incorrect or incomplete physics in the models, or systematic issues that arise in the discretization of the model grids.

There are currently around 20 groups around the world that develop such models and which contribute to the standardized archives that have been developed and made available to outside researchers. The Coupled Model Intercomparison Project version 3 (CMIP3) archive was initially developed to support the IPCC 4th Assessment Report (published in 2007) [37], but has subsequently been used in over 500 publications and continues to be a rich source of climate simulation output.

2.2. **Related work in Climate Science.** The model projections for many aspects of climate change are robust for some quantities (regional temperature trends for instance), but vary significantly across different models for other equally important metrics (such as regional precipitation). Given those uncertainties, climate researchers have looked for simple ways to judge model skill so that projections can be restricted (or weighted towards) models with more skill [16, 17, 35]. Any attempt at model ranking or weighting must include justification that the choices are meaningful for the specific context. One approach is to make a "perfect model" assumption (i.e. that one model is the "truth") and then track whether a methodology trained on the "true" model over a calibration interval can continue to skillfully track that simulation in the forecast period. Work on this problem and related discussions was recently the subject of an IPCC Expert Meeting on Assessing and Combining Multi-Model Climate Projections, where we presented our preliminary results [27].

A number of studies have looked at how the multi-model ensemble can be used to enhance information over and above the information available from just one model. For instance, the simple average of the models' output gives a better estimate of the real world than any single model [32, 33]. This is surprising because the models are not a random selection from a space of all possible climate models, but rather an interdependent ensemble. Indeed, the reduction in root mean square errors plateaus after about 10 models are included in the average and does not follow the $1/\sqrt{n}$ path one would expect for truly random errors. This behaviour can be expected if the individual models are statistically indistinguishable from the "truth," rather than an independent estimate of the truth plus some error [4]. Finally, more sophisticated ensemble methods are being explored, for instance in the case of regional climate models (see e.g. [34] and references therein).

2.3. **Tracking climate models.** Given the current assumption that the multi-model mean is the best estimate of climatology, it has often been implicitly assumed that the multi-model ensemble mean is also the best projection for the future. However, while this has not been demonstrated in either practice or theory, it has nonetheless become the default strategy adopted by IPCC and other authors. Other approaches have been tried (using skill measures to create weights among the models, creating emulators from the model output that map observables to projections), but rigorous support for these approaches, or even a demonstration that they make much difference, has so far been patchy.

In this work, we use machine learning on hindcasts from the CMIP3 archive and over 100 years of observed global mean temperature anomalies, to demonstrate an algorithm that tracks the changing sequence of which model currently predicts best. A *hindcast* is a model simulation of a past period for which we have a good idea how the external drivers changed; it is not a replication of the specific weather that occurred. Our algorithm attains lower mean prediction loss than predicting with the average over model predictions. This is an impactful result because to date, the average of all models' predictions was believed to be the best single predictor of the whole sequence [32, 33]. We also demonstrate the utility of the algorithm when trained on future climate model projections, using any one model's predictions to simulate the observations.

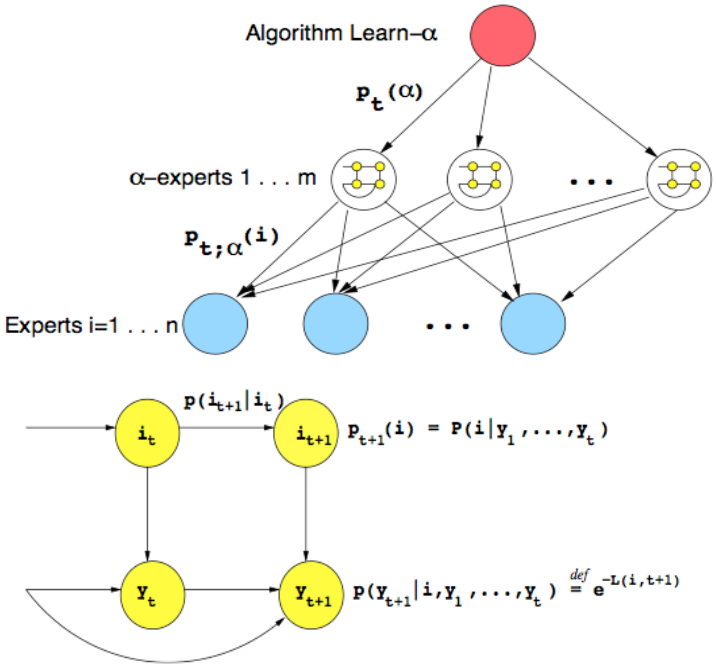

FIGURE 2. Top figure: a. The Learn-α algorithm of [25]. The α-experts are Fixed-Share(α) algorithms from [15]. Bottom figure: b. The generalized Hidden Markov Model corresponding to the algorithms of [15].

3. ALGORITHMS

We apply the Learn-α algorithm of Monteleoni and Jaakkola [25] to track a shifting sequence of temperature values with respect to the predictions of "experts," instantiated as climate models. This is an *online learning* algorithm, which is useful in this setting because the eventual goal is to make both real-time and future predictions. A large class of online learning algorithms have been designed for the framework in which no statistical assumptions are made about the sequence of observations, and algorithms are evaluated based on *regret*: relative prediction loss with respect to the hindsight-optimal algorithm in a comparator class (e.g. [22, 15]; there is a large literature, see [8] for a thorough treatment). Many such algorithms, designed for predicting in non-stationary environments, descend from variants of an algorithm due to Herbster and Warmuth [15], which is a form of multiplicative update algorithm. Their Fixed-Share algorithm tracks a sequence of observations with respect to a set of n experts' predictions, by updating a probability distribution $p_t(i)$ over experts, i, based on their current performance, and making predictions as a function of the experts' predictions, subject to this distribution. The authors proved performance guarantees for this algorithm with respect to the best k-segmentation of a finite sequence of observations into k variable-length segments, and assignment of the best expert per segment.

As illustrated in [25], this class of algorithms can be derived as Bayesian updates of an appropriately defined Hidden Markov Model (HMM), where the current best expert is the hidden variable. (Despite the Bayesian re-derivation, the regret analyses require no assumptions on the observations.) As shown in Figure 2b, equating the prediction loss function (for the given problem) to the negative log-likelihood of the observation given the expert, yields a (generalized) HMM, for which Bayesian updates correspond to the weight updates in the Fixed-Share algorithm, when the transition matrix

Algorithm Learn-α for Tracking Climate Models
Input:
Set of climate models, M_i, $i \in \{1, \cdots, n\}$ that output predictions $M_i(t)$ at each time t.
Set of $\alpha_j \in [0,1]$, $j \in \{1, \cdots, m\}$: discretization of α parameter.
Initialization:
$\forall j,\ p_1(j) \leftarrow \frac{1}{m}$
$\forall i, j,\ p_{1,j}(i) \leftarrow \frac{1}{n}$
Upon tth data observation, y_t:
For each $i \in \{1 \dots n\}$:
Loss$[i] \leftarrow (y_t - M_i(t))^2$
For each $j \in \{1 \dots m\}$:
LossPerAlpha$[j] \leftarrow -\log \sum_{i=1}^{n} p_{t,j}(i)\, e^{-\text{Loss}[i]}$
$p_{t+1}(j) \leftarrow p_t(j) e^{-\text{LossPerAlpha}[j]}$
For each $i \in \{1 \dots n\}$:
$p_{t+1,j}(i) \leftarrow \sum_{k=1}^{n} p_{t,j}(k)\, e^{-\text{Loss}[k]}\, P(i\|k;\alpha_j)$
Normalize $P_{t+1,j}$
PredictionPerAlpha$[j] \leftarrow \sum_{i=1}^{n} p_{t+1,j}(i)\, M_i(t+1)$
Normalize P_{t+1}
Prediction $\leftarrow \sum_{j=1}^{m} p_{t+1}(j)$ PredictionPerAlpha$[j]$

FIGURE 3. Algorithm Learn-α, due to [25], applied to tracking climate models.

is simply $(1-\alpha)$ for self-transitions, and $\alpha/(n-1)$ for transitions to any of the other $(n-1)$ experts. The parameter $\alpha \in [0,1]$ models how likely switches are to occur between best experts.

In [25, 29] it was shown theoretically and empirically that the wrong setting of α for the sequence in question can lead to poor performance. The authors derived upper and lower regret bounds (with respect to Fixed-Share using the hindsight-optimal α) for this class of online learning algorithms. They provided an algorithm, Learn-α, that learns this parameter online, simultaneous to performing the original learning task, and showed that it avoids the lower bound and yields better performance guarantees: regret is logarithmic, as opposed to linear, in the number of predictions. Learn-α uses a hierarchical model shown in Figure 2a, with a set of meta-experts: sub-algorithms that are instances of Fixed-Share. Each sub-algorithm of Learn-α runs Fixed-Share(α_j), where α_j, $j \in \{1, \cdots, m\}$, forms a discretization of the α parameter. At the top of the hierarchy, the algorithm learns the parameter α, by tracking the meta-experts. In order to learn the best fixed value of α, a similar model is used, with self-transition probabilities of 1.

Figure 3 shows our application of the algorithm Learn-α to the problem of tracking climate models. The experts are instantiated as the climate models; each model produces one prediction per unit of time, and we denote the true observation at time t, by y_t. The algorithm is modular with respect to loss function; we chose squared loss since it is a simple loss, useful in regression problems.

Regret-optimal parameter discretization. We use a discretization procedure for the pareameter α given in [25] which optimizes the regret bound. The input to the procedure is T, the desired number of iterations of online learning. Since the regret-optimal discretization is a function of T, we use a different set of α values for past data than for model prediction data that starts in the past and continues into the future. Recent work has further studied the issues of discretizing an analogous parameter for similar algorithms [13].

4. Data and Experiments

4.1. **Data.** We ran experiments with our application of the Learn-α algorithm on historical temperature data from 1900 through 2008 as well as the corresponding predictions of 20 different climate models, per year. It is important to emphasize that climate models are not data-driven models

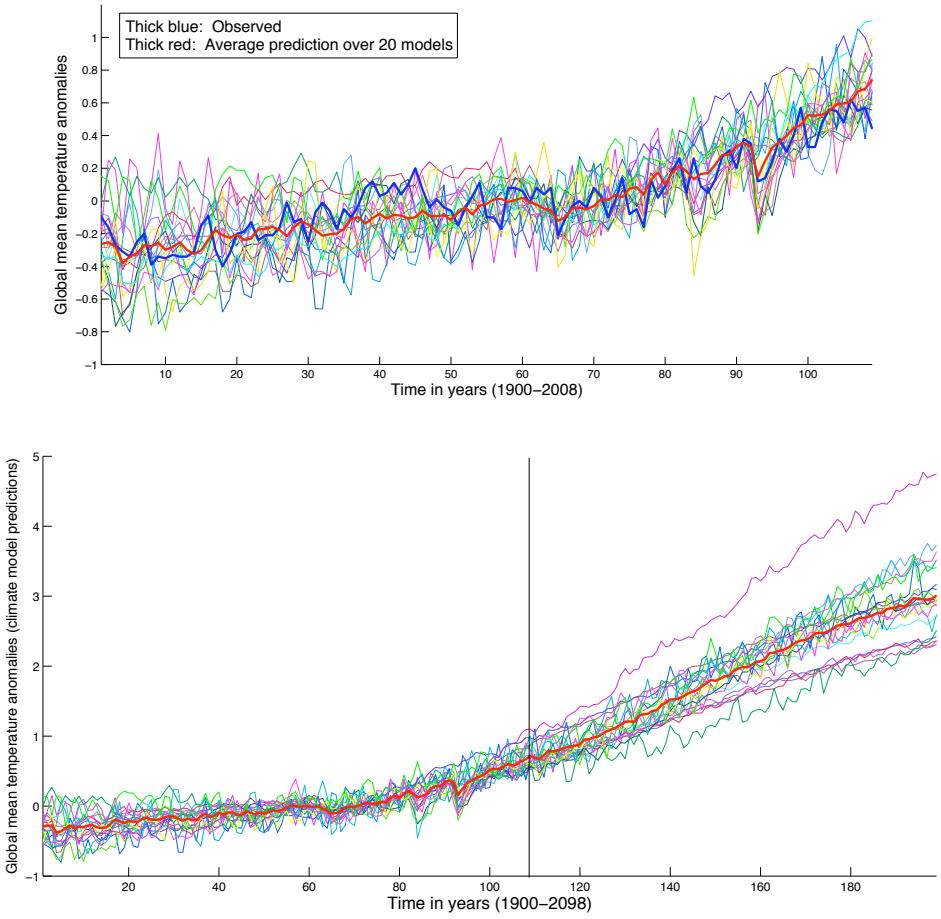

FIGURE 4. Top figure: a. Observations and model predictions through 2008. Bottom figure: b. Model predictions through 2098. The black vertical line separates past from future.

but rather complex mathematical models based on geophysical and meteorological principles. In particular they are not "trained" on data as is done with machine learning models. Therefore it is valid to run them predictively on past data.

Both the climate model predictions, and the true observations, are in the form of global mean temperature anomalies. (The model predictions are from the CMIP3 archive [2], and the temperature anomalies are available from NASA [3].) A *temperature anomaly* is defined as the difference between the observed temperature and the temperature at the same location at a fixed, benchmark time. Anomalies are therefore measurements of change in temperature. When studying global mean temperature, it is useful to use anomalies, because, while temperatures vary widely over geographical location, temperature anomalies typically vary less. For example, at a particular time it might be 80°F in New York, and 70°F in San Diego, but the anomaly from the benchmark time might be 1°F in both places. Thus there is lower variance when temperatures anomalies are averaged over many geographic locations, than when using temperatures. The data we use has been averaged over many geographical locations, and many times in a year, yielding one value for global mean temperature anomaly per year. (In this case the benchmark is averaged over 1951-80; one can convert between

benchmark eras by subtracting a constant.) Figure 4 shows the model predictions, where the thick red line is the mean prediction over all models, in both plots. The thick blue line indicates the true observations.

We also ran experiments using climate model projections into the 21st century, as we had model predictions through 2098. In this case, we used any one model's predictions as the quantity to learn, based only on the predictions of the remaining 19 models. The motivation for the future simulation experiments are as follows. Future climates are of interest, yet there is no observation data in the future, with which to evaluate machine learning algorithms. Furthermore, given the significant fan-out that occurs among model predictions starting after 2009 and increasing into the future (see Figure 4b), it may no longer make sense to predict with the mean prediction; that is, the average prediction diverges over time from most individual model predictions. However, we do want to be able to harness the predictions of the climate models in forming our future predictions. Given these reasons, and the climate science community's interest in the "perfect model" assumption, we evaluated algorithms on predicting the labels generated by one climate model, using the remaining models as input.

Further data details. While some models produced predictions slightly earlier than 1900, this was not the case with all models. The earliest year at which we had predictions from all 20 models was 1900. Some climate models have only one simulation run available in the data, while others have up to 7. We obtained similar results to those we report below by training on the average over runs of each model, however climate scientists do not view that scenario as an actual simulation. Thus we arbitrarily picked one run per model, for each of the 20 models, as input to all the algorithms.

The climate models contributing to the CMIP3 archive include those from the following laboratories: Bjerknes Center for Climate Research (Norway), Canadian Centre for Climate Modelling and Analysis, Centre National de Recherches Météorologiques (France), Commonwealth Scientific and Industrial Research Organisation (Australia), Geophysical Fluid Dynamics Laboratory (Princeton University), Goddard Institute for Spaces Studies (NASA), Hadley Centre for Climate Change (United Kingdom Meteorology Office), Institute of Atmospheric Physics (Chinese Academy of Sciences), Istituto Nazionale di Geofisica e Vulcanologia (Italy), Institute of Numerical Mathematics Climate Model (Russian Academy of Sciences), Model for Interdisciplinary Research on Climate (Japan), Meteorological Institute at the University of Bonn (Germany), Max Planck Institute (Germany), Meteorological Research Institute (Japan), National Center for Atmospheric Research (Colorado), among others.

4.2. **Experiments and results.** In addition to Learn-α, we also experimented with the following algorithms: simply predicting with the mean prediction over the experts, doing so with the median prediction, and performing batch linear regression (least squares) on all the data seen so far. The regression problem is framed by considering the vector of expert predictions at a given year as the example, and the true observation for that year as the label. Batch linear regression has access to the entire past history of examples and labels.

The four future simulations reported use labels from 1) `giss model e r run4`, 2) `mri cgcm2 3 2a run5`, 3) `ncar ccsm3 0 run9`, 4) `cnrm cm3 run1`. The labeling runs for the future simulations were chosen (over all runs of all models) to represent the range in past performance with respect to average prediction loss. 1) is the best performing model, 4) is the worst, 3) attains the median, and 2) performs between 1) and 3), at the median of that range. For each simulation, the remaining 19 climate models' predictions are used as input.

In Table 1, we compare mean loss on real-time predictions, i.e. predictions per year, of the algorithms. This is a standard evaluation technique for online learning algorithms. Several of the algorithms are online, including Learn-α and the techniques of simply forming predictions as either the mean or the median of the climate models' predictions. (For the future simulations, the annual mean and median predictions are computed over the 19 climate models used as input.) Least squares linear regression operates in a batch setting, and cannot even compute a prediction unless

Algorithm:	Historical	Future Sim. 1	Future Sim. 2	Future Sim. 3	Future Sim. 4
Learn-α Algorithm	0.0119	0.0085	**0.0125**	**0.0252**	**0.0401**
	$\sigma = 0.0002$	$\sigma = 0.0001$	$\sigma = 0.0004$	$\sigma = 0.0010$	$\sigma = 0.0024$
Linear Regression*	0.0158	**0.0051**	0.0144	0.0264	0.0498
	$\sigma = 0.0005$	$\sigma = 0.0001$	$\sigma = 0.0004$	$\sigma = 0.0125$	$\sigma = 0.0054$
Best Expert	**0.0112**	0.0115	0.0286	0.0301	0.0559
	$\sigma = 0.0002$	$\sigma = 0.0002$	$\sigma = 0.0014$	$\sigma = 0.0018$	$\sigma = 0.0053$
Average Prediction	0.0132	0.0700	0.0306	0.0623	0.0497
	$\sigma = 0.0003$	$\sigma = 0.0110$	$\sigma = 0.0016$	$\sigma = 0.0055$	$\sigma = 0.0036$
Median Prediction	0.0136	0.0689	0.0308	0.0677	0.0527
	$\sigma = 0.0003$	$\sigma = 0.0111$	$\sigma = 0.0017$	$\sigma = 0.0070$	$\sigma = 0.0038$
Worst Expert	0.0726	1.0153	0.8109	0.3958	0.5004
	$\sigma = 0.0068$	$\sigma = 2.3587$	$\sigma = 1.4109$	$\sigma = 0.5612$	$\sigma = 0.5988$

TABLE 1. Mean and variance of annual losses. The best score per experiment is highlighted. *Linear Regression cannot form predictions for the first 20 years (19 in the future simulations), so its mean is over fewer years than all the other algorithms.

the number of examples it trains on is at least the dimensionality, which in this case is the number of experts. We also compare to the loss of the best and worst expert. Computing the identity of "best" and "worst," with respect to their prediction losses on the sequence, can only be done in hindsight, and thus also requires batch access to the data. (For the future simulations, the identity of the best and worst at predicting the labels generated by one climate model is determined from the remaining 19 climate models). We test batch linear regression using this method as well, computing its error in predicting just the current example, based on all past data. Note that although all examples are used for training, they also contribute to error, before the label is viewed, so this online learning evaluation measure is comparable (but not identical) to a form of test error (in the batch setting). In particular, this "progressive validation" error was analyzed in [5], which provided formal bounds relating it, as well as k-fold cross-validation error, to standard batch holdout error, in certain settings.

Learn-α's peformance, with respect to the average over all model predictions, is a break-through; as that was the current state-of-the-art. As shown in Table 1, in every experiment, Learn-α suffers lower mean annual loss than predicting using the average over all model predictions. Furthermore, Learn-α surpasses the performance of the best expert in all but one experiment (Historical), in which its performance nearly matches it. Similarly, Learn-α surpasses the performance of least squares linear regression in all but one experiment (Future Simulation 1), in which its performance is still close. Learn-α's outperformance of batch linear regression on almost all experiments suggests that weighting all historical data equally (as does linear regression) produces worse predictions of the present observation, than using a weighting that focuses more on the recent past (as Learn-α does implicitly). This helps lend validity to the use of online learning algorithms in the climate change prediction domain.

Remark. An interesting result is that on historical data, the best climate model outperforms the average prediction over climate models. This appears to contradict the related work in climate science [32, 33]. Reichler and Kim [32] were concerned with performance dominance across multiple metrics, as opposed to just prediction loss on global mean temperature anomalies, and thus there is no contradiction. Reifen and Toumi [33] consider model prediction runs from the same archive as we do, however their experimental set-up differs. Predictions from 17 models are evaluated through 1999, with respect to a different set of observation data. Regardless of the finding that in our setting there is a model that performs better than the average, the "best" expert cannot be used as a

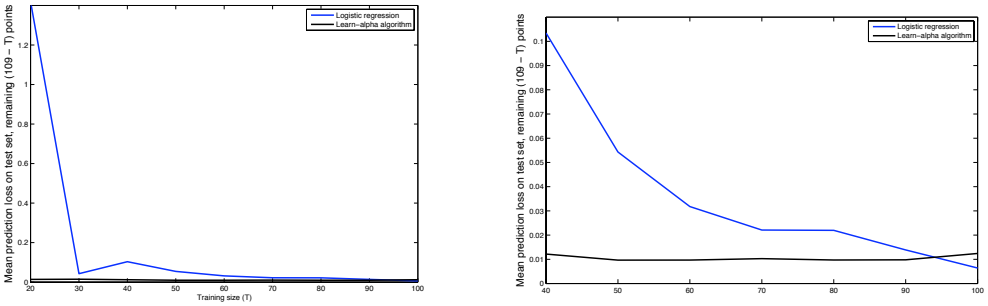

FIGURE 5. Batch evaluations. Plot of mean test error on the remaining points, when only the first T are used for training. Right plot zooms in on T $\geq$ 40 (x-axis).

prediction technique in practice, since knowledge of which model performs best requires observation of the entire data set, a scenario that is impossible in a future prediction problem.

4.3. **Batch comparison of the learning algorithms.** Since least squares linear regression is a batch algorithm, here we provide a batch-like comparison of the two machine learning algorithms. Because this data is measured over time, there is importance in its ordering, and thus it is not appropriate to use standard cross-validation with multiple folds. Instead we use the first part of the data as the training data, and the remaining data for testing, for various values of the split location, from 20 to 100. We chose this range for the possible splits because least squares linear regression needs at least the number of training points as the dimensionality (20 in this case, the number of climate models), in order to compute a classifier, and there are only 109 years of historical data.

Figure 5 shows that for most values of the split between training data and test data, Learn-α suffers lower mean test error. The one split on which this does not hold (100), contains only 9 points in the test set, so both measurements have high variance; indeed the difference in mean test error at $T = 100$ is less than one standard deviation of Learn-α's test error ($\sigma = 0.0185$). These results suggest that the non-stationary nature of the data, coupled with the limited amount of historical data, poses challenges to a naïve batch algorithm. Just as the results in Table 1 suggest that weighting all historical data equally produces worse predictions of the present observation than a weighting that focuses more on the recent past, in this batch-like evaluation setting, Figure 5 reveals that a similar conclusion also holds for predictions into the future. That is, as far as annual global mean temperature anomalies are concerned, the present (or recent past) appears to be a better predictor of the future than the past.

4.4. **Learning curves.** Here we provide learning curves for Learn-α, plotted against the best and worst experts in hindsight, and the average over expert predictions, which was the previous benchmark. These experiments generated the statistics summarized in Table 1. Figure 6 plots the squared error between predicted and observed annual mean temperature, by year from 1900 to 2008. Learn-α suffers less loss than the mean over model predictions on over 75% of the years (82/109).

The learning curves from the future simulation experiments, Figures 7-8, demonstrate that Learn-α is very successful at predicting one model's predictions for future predictions up to the year 2098. This is notable, as the future projections vary widely among the climate models. In each of the four future simulations, the (blue) curve indicating the worst model (with respect to predicting the model in question) varies increasingly into the future, whereas our algorithm (black) tracks, and in fact surpasses, the performance of the best model (green). Including these simulations, in 10 future simulations that we ran, each with a different climate model providing the labels, Learn-α suffers less loss than the mean over the remaining model predictions on, 75%-90% of the years.

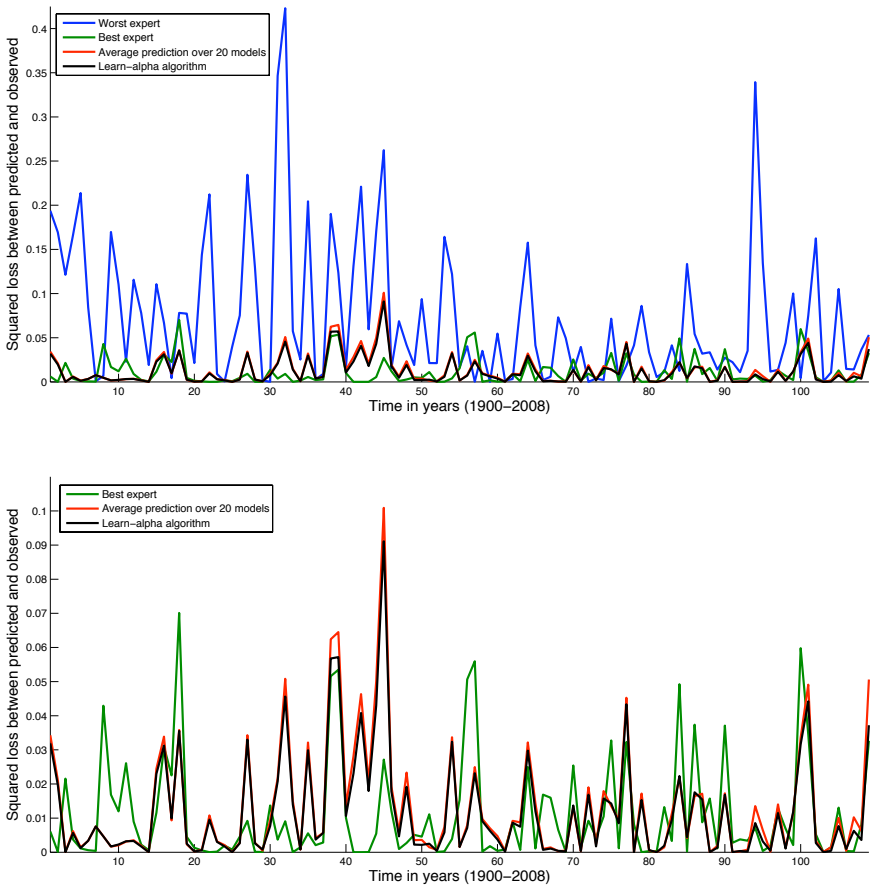

FIGURE 6. Squared loss between predicted and observed global mean temperature anomalies. The bottom plot zooms in on the y-axis.

4.5. **Weight evolution.** We also provide plots of the evolution of the weights on climate models, and internal sub-algorithms, as they were learned by Learn-α in the historical data experiment.

Figure 9a illustrates how the Learn-α algorithm updates weights over the sub-algorithms, instances of the Fixed-Share(α) algorithm running with different values of α. The Learn-α algorithm tracks the best *fixed* value of the α parameter, so as the plot shows, one alpha consistently receives an increasing fraction of the weight. The α value that received the highest weight at the end was the smallest, which was 0.0046 for the historical data experiments.

Figure 9b illustrates how a Fixed-Share sub-algorithm (in this case $\alpha = 0.0046$) updates weights over the climate models. The algorithm predicts with a linear combination of the climate model predictions. As opposed to tracking the best *fixed* climate model, or linear combination, the linear combination of climate models changes dynamically based on the currently observed performance of the different climate models. The climate model which received the highest weight at the end was `giss model e r run4`, which is also the best performing expert on the historical data set.

5. Discussion and Future Work

The exciting challenge begged by our encouraging results, is how to track climate models when predicting *future* climates. The current state-of-the-art tracking methods still rely on receiving true

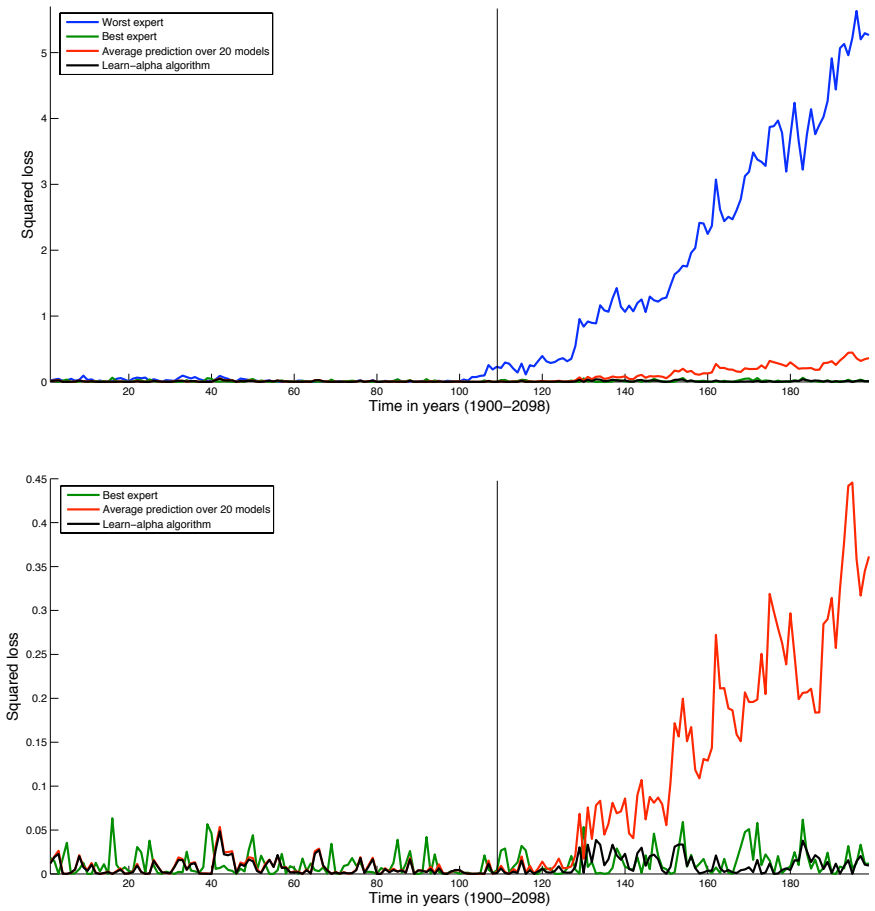

FIGURE 7. Future Simulation 1: Tracking the predictions of one model using the predictions of the remaining 19 as input, with no true temperature observations. Black vertical line separates past from future. Bottom plot zooms in on y-axis.

observations, with which to evaluate the models' predictions. Our goal is to design algorithms that can track models in unsupervised, or semi-supervised settings. The analysis poses challenges however; providing (standard) regret bounds for the fully unsupervised setting is likely impossible, and we are not aware of any related work. We can also consider a *semi-supervised learning* setting [10]. There is some literature on regret analyses of semi-supervised online learning; [9, 7] consider the special case of active learning. Another related setting is that of imperfect monitoring, in which the learner has access to partial feedback, but not the true observations, e.g. [24]. One approach that we have shown to be feasible in practice (see Figures 7-8), is to view expert predictions themselves as partial feedback, in order to design semi-supervised algorithms. We can also turn to the batch setting, when one-time predictions are needed, given past data. However our preliminary experiments with batch linear regression do not surpass the performance of our online technique. Noting that predictions are sometimes only requested for certain benchmark years, (e.g. 2020, 2050, 2100), it may be worth considering a transductive model, and experimenting with methods for transductive regression [11, 12].

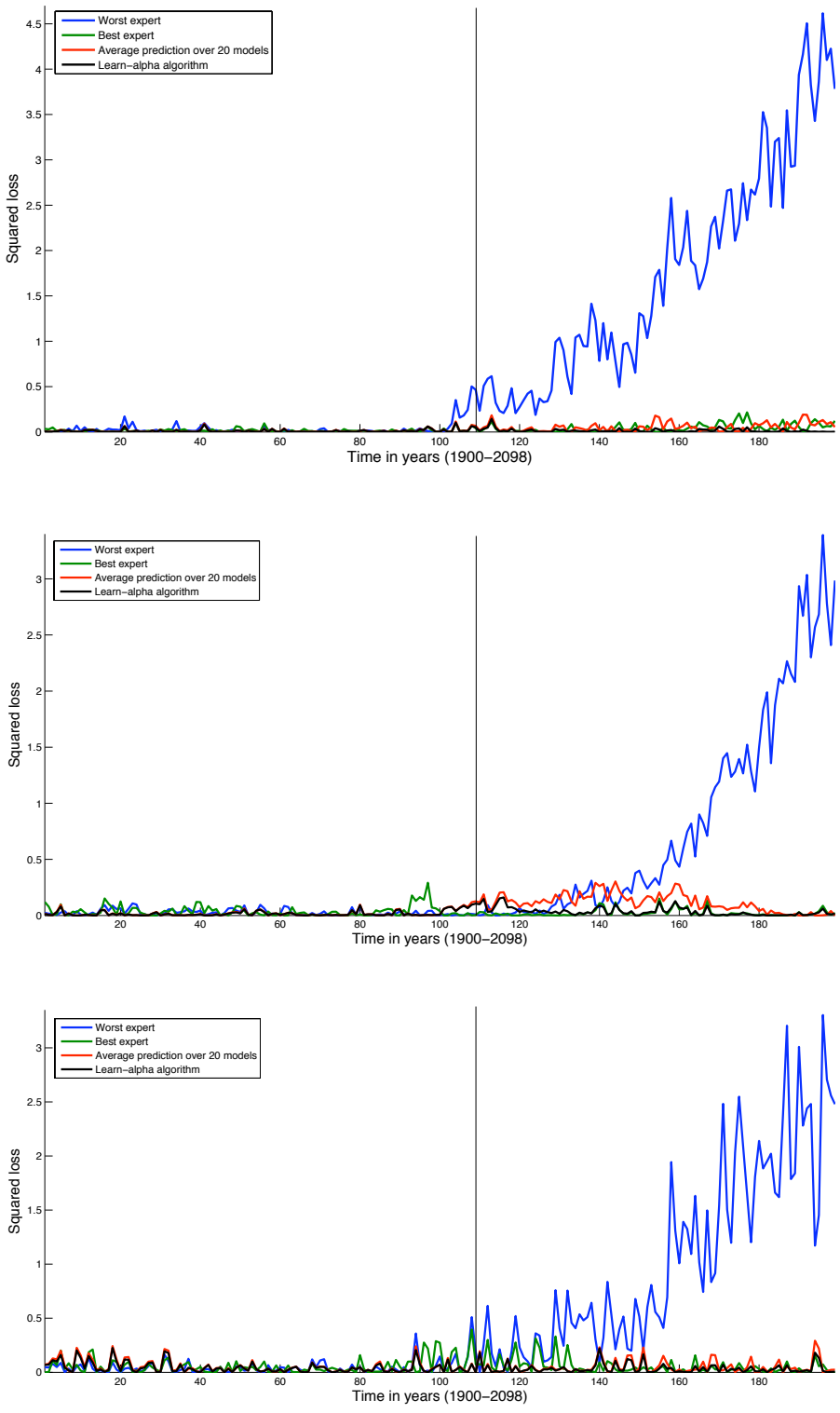

FIGURE 8. Top: Future Sim 2, Middle: Future Sim. 3, Bottom: Future Sim. 4.

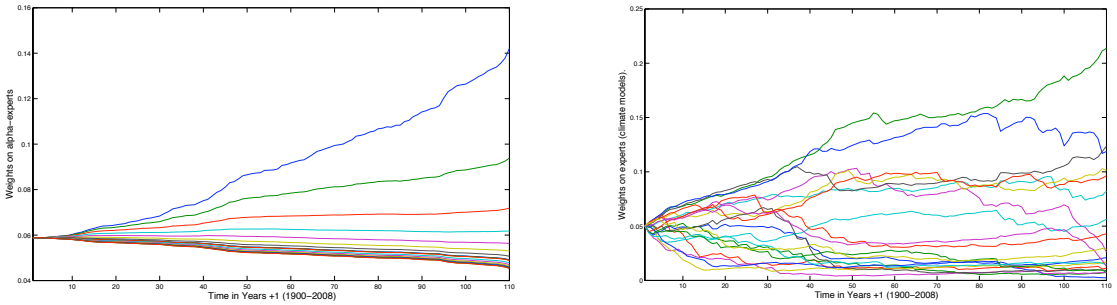

FIGURE 9. Weight evolution. Top figure: a. Algorithm's weights on α-experts. Bottom figure: b. Best α-expert's weights on experts (climate models).

In summary, our results advance the state-of-the-art in the climate science community, with respect to combining climate model predictions. Our methods are applicable to any quantity predicted by a set of climate models, and we plan to use them for predicting at smaller regional scales, and shorter times scales, as well as predicting other important climate benchmarks, such as carbon dioxide. In addition to our specific contributions, we hope to inspire future applications of machine learning to improve climate predictions and to help answer pressing questions in climate science.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for CIDU 2010, and for the Temporal Segmentation Workshop at NIPS 2009, as well as the anonymous reviewers and the participants of The Learning Workshop (Snowbird) 2010, especially Yann LeCun, for helpful comments on earlier versions of this work.

REFERENCES

[1] `http://celebrating200years.noaa.gov/breakthroughs/climate_model/welcome.html`.
[2] `http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php`.
[3] `http://data.giss.nasa.gov/gistemp/`.
[4] J. D. Annan and J. C. Hargreaves. Reliability of the CMIP3 ensemble. *Geophys. Res. Lett.*, page L02703, 2010.
[5] A. Blum, A. Kalai, and J. Langford. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT '99)*, pages 203–208, 1999.
[6] A. Braverman, R. Pincus, and C. Batstone. Data mining for climate model improvement. In *Sixth Annual NASA Earth Science Technology Conference*, 2006.
[7] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Worst-case analysis of selective sampling for linear-threshold algorithms. *Journal of Machine Learning Research*, 7:1205–1230, 2006.
[8] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
[9] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005.
[10] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.
[11] C. Cortes and M. Mohri. On transductive regression. In *Advances in Neural Information Processing Systems 21*, pages 305–312, 2007.
[12] C. Cortes, M. Mohri, D. Pechyony, and A. Rastogi. Stability of transductive regression algorithms. In *Proceedings of the Twenty-fifth International Conference on Machine Learning*, 2008.
[13] S. de Rooij and T. van Erven. Learning the switching rate by discretising bernoulli sources online. In *AISTATS '09: Proc. Twelfth International Conference on Artificial Intelligence and Statistics*, 2009.
[14] E. Hawkins and R. Sutton. The potential to narrow uncertainty in regional climate predictions. *Bull. Amer. Meteor. Soc.*, 90:1095–1107, 2009.
[15] M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.
[16] R. Knutti. The end of model democracy? *Climatic Change*, page (in press), 2010.
[17] R. Knutti, J. C. R. Furrer, C. Tebaldi, and G. A. Meehl. Challenges in combining projections from multiple climate models. *J. Climate*, page (in press), 2010.

[18] V. Krasnopolsky, M. Fox-Rabinovitz, and A. Belochitski. Decadal Climate Simulations Using Accurate and Fast Neural Network Emulation of Full, Longwave and Shortwave, Radiation. *Monthly Weather Review*, 136:368–3695, 2008.
[19] V. M. Krasnopolsky and M. S. Fox-Rabinovitz. Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. *Neural Networks*, 19(2):122–134, 2006.
[20] V. Kumar. Discovery of patterns in global earth science data using data mining. In *PAKDD (1)*, 2010.
[21] C. Lima, U. Lall, T. Jebara, and A. Barnston. Statistical prediction of enso from subsurface sea temperature using a nonlinear dimensionality reduction. *Journal of Climate*, 22(17):4501–4519, 1 September 2009.
[22] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 256–261, 1989.
[23] A. C. Lozano, H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. R. M. Hosking, and N. Abe. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 587–596, 2009.
[24] G. Lugosi, S. Mannor, and G. Stoltz. Strategies for prediction under imperfect monitoring. In *Proc. 20th Annual Conference on Learning Theory*, 2007.
[25] C. Monteleoni and T. Jaakkola. Online learning of non-stationary sequences. In *NIPS '03: Advances in Neural Information Processing Systems 16*, 2003.
[26] C. Monteleoni, S. Saroha, and G. Schmidt. Tracking climate models. In *Temporal Segmentation Workshop, at the Conference on Neural Information Processing Systems*, 2009.
[27] C. Monteleoni, S. Saroha, and G. Schmidt. Can machine learning techniques improve forecasts? In *Intergovernmental Panel on Climate Change (IPCC) Expert Meeting on Assessing and Combining Multi Model Climate Projections*, 2010.
[28] C. Monteleoni, S. Saroha, and G. Schmidt. Tracking climate models. In *The Learning Workshop, Snowbird*, 2010.
[29] C. E. Monteleoni. Online learning of non-stationary sequences. SM Thesis. In *MIT Artificial Intelligence Technical Report 2003-011*, 2003.
[30] D. R. Musicant, J. M. Christensen, and J. F. Olson. Supervised learning by training on aggregate outputs. In *Proceedings of the Seventh IEEE International Conference on Data Mining*, pages 252–261, 2007.
[31] R. Ramakrishnan, J. J. Schauer, L. Chen, Z. Huang, M. Shafer, D. S. Gross, and D. R. Musicant. The EDAM project: Mining atmospheric aerosol datasets. *International Journal of Intelligent Systems*, 20(7):759–787, 2005.
[32] T. Reichler and J. Kim. How well do coupled models simulate today's climate? *Bull. Amer. Meteor. Soc.*, 89:303–311, 2008.
[33] C. Reifen and R. Toumi. Climate projections: Past performance no guarantee of future skill? *Geophys. Res. Lett.*, 36, 2009.
[34] S. Sain and R. Furrer. Combining climate model output via model correlations. *Stochastic Environmental Research and Risk Assessment*, 2010.
[35] B. D. Santer, K. E. Taylor, P. J. Gleckler, C. Bonfils, T. P. Barnett, D. W. Pierce, T. M. L. Wigley, C. Mears, F. J. Wentz, W. Brueggemann, N. P. Gillett, S. A. Klein, S. Solomon, P. A. Stott, and M. F. Wehner. Incorporating model quality information in climate change detection and attribution studies. *Proc. Nat. Acad. Sci.*, 106:14,778–14783, 2009.
[36] Schmidt, G.A., R. Ruedy, J. Hansen, I. Aleinov, N. Bell, M. Bauer, S. Bauer, B. Cairns, V. Canuto, Y. Cheng, A. D. Genio, G. Faluvegi, A. Friend, T. Hall, Y. Hu, M. Kelley, N. Kiang, D. Koch, A. Lacis, J. Lerner, K. Lo, R. Miller, L. Nazarenko, V. Oinas, J. Perlwitz, J. Perlwitz, D. Rind, A. Romanou, G. Russell, M. Sato, D. Shindell, P. Stone, S. Sun, N. Tausnev, D. Thresher, and M.-S. Yao. Present day atmospheric simulations using GISS ModelE: comparison to in-situ, satellite and reanalysis data. *Journal of Climate*, 19:153–192, 2006.
[37] S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, editors. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.
[38] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter. Discovery of climate indices using clustering. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 446–455, 2003.
[39] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly. An exploration of climate data using complex networks. *ACM SIGKDD Explorations*, 12(1), (to appear) 2010.