Tracking Climate Models

Claire Monteleoni[1]*, Gavin A. Schmidt[2], Shailesh Saroha[3] and Eva Asplund[3,4]

[1] *Center for Computational Learning Systems, Columbia University, New York, NY, USA*

[2] *Center for Climate Systems Research, NASA Goddard Institute for Space Studies, Columbia University, New York, NY, USA*

[3] *Department of Computer Science, Columbia University, New York, NY, USA*

[4] *Barnard College, Columbia University, New York, NY, USA*

Received 1 December 2010; revised 16 April 2011; accepted 21 April 2011
DOI:10.1002/sam.10126
Published online 8 July 2011 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Climate models are complex mathematical models designed by meteorologists, geophysicists, and climate scientists, and run as computer simulations, to predict climate. There is currently high variance among the predictions of 20 global climate models, from various laboratories around the world, that inform the Intergovernmental Panel on Climate Change (IPCC). Given temperature predictions from 20 IPCC global climate models, and over 100 years of historical temperature data, we track the changing sequence of which model predicts best at any given time. 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, 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 global mean temperature 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 is the default practice 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. To complement our global results, we also ran experiments on IPCC global climate model temperature predictions for the specific geographic regions of Africa, Europe, and North America. On historical data, at both annual and monthly time-scales, and in future simulations, our algorithm typically outperformed both the best climate model per region and linear regression. Notably, our algorithm consistently outperformed the average prediction over models, the current benchmark. © 2011 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 4: 372–392, 2011

Keywords: machine learning; climate science; climate modeling; online learning; prediction with expert advice; data mining

1. INTRODUCTION

The threat of climate change is one of the greatest challenges currently facing society. 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, ocean circulation, and river flow. 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

Correspondence to: Claire Monteleoni (cmontel@ccls.columbia.edu)

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 ever more 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, due in 2013.

Currently there is very high variance among the predictions of these 20 models, even for *identical* future scenarios. 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. While the variance is high however, 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 [1,2].

Our contribution is an application of a machine learning algorithm that produces predictions that match or surpass that of the best climate 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 suggest 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 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 [1,2].

1.1. 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 [3,4], analyzing the impacts of climate change [5], and calibrating a climate model [6]. Clustering techniques have been developed to model climate data [7]. Machine learning has been applied to predicting the El Niño climate pattern [8], and modeling climate data [9]. In another work, machine learning and data mining researchers proposed the use of data-driven techniques for climate change attribution [10]. There has also been work on integrating neural networks into global climate models [11,12]. In the field of weather prediction, which is concerned with predicting at much shorter time-scales than those studied in climate science, machine learning techniques have enjoyed success in practice, e.g. [13].

We are not aware of applications, beyond our own, of machine learning to the problem of tracking global climate models. We apply the Learn-α algorithm of Monteleoni and Jaakkola [14] 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 Fig. 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 [17]. It is important to note that unlike the use of the term *model* in machine learning, here it denotes systems of mathematical models, that are *not* data-driven. These complex systems are composed of individual mathematical models of each of the processes mentioned, as well as many other processes. 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

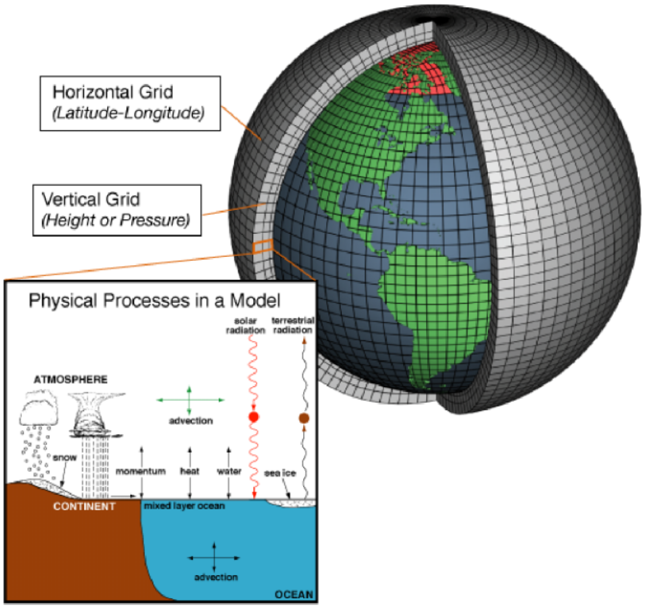

Fig. 1 Global climate model (schematic due to [16]). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

drivers of climate (such as human-caused increases in greenhouse gases and aerosols, large volcanic eruptions, solar activity, 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 can be uncertain, in both the past and the future. 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 "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 [18].

For projections of future climate, there are three separate components to the uncertainty [19]. 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.

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 World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model data set archive was initially developed to support the IPCC 4th Assessment Report (published in 2007) [20], 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 toward) models with more skill [18,21,22]. 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

[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.

discussions was recently the subject of an IPCC Expert Meeting on Assessing and Combining Multi-Model Climate Projections, at which a preliminary version of this work appeared [23].

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 [1,2]. This is surprising because the models are not independent in a statistical sense. That is, they are not a random sample from the space of all possible climate models, but rather an interdependent ensemble. Therefore the law of large numbers need not apply (i.e. increasing the number of models in an average need not approach the "truth"). Indeed, the reduction in root mean square errors plateaus after about ten models are included in the average and does not follow the $1/\sqrt{n}$ path one would expect for truly random errors. Although one cannot assume that the models are all clustered around "truth", recent approaches in climate science consider individual models and the "truth" as being drawn from the same distribution (e.g. [24,25]). There are also other lines of work on how to quantify the confidence on climate model projections, and ensembles thereof, e.g. [22,26,27].

Finally, there has been recent work on developing and applying more sophisticated ensemble methods [28–35]. For example, Smith *et al.* [34] propose uni- and multivariate Bayesian approaches to combine the predictions over a variety of locations of a multi-model ensemble, in the batch setting. In the case of regional climate models, Sain and Furrer [35] propose ensemble methods involving multivariate Markov random fields.

2.3. Tracking Climate Models

Given that the multi-model mean is the current best estimate of climatology, it has often been implicitly assumed that the multi-model ensemble mean is also the best projection for the future. 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 temperature data, 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 relatively good idea how the external drivers changed; it is not a replication of the specific weather that occurred. In a variety of experimental scenarios, at both global and regional scales, 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 [1,2]. 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.

3. ALGORITHMS

We apply the Learn-α algorithm of Monteleoni and Jaakkola [14] 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. [15,36]; there is a large literature, see [37] 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 the work of Monteleoni and Jaakkola [14], 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 panel (a) of Fig. 2, 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

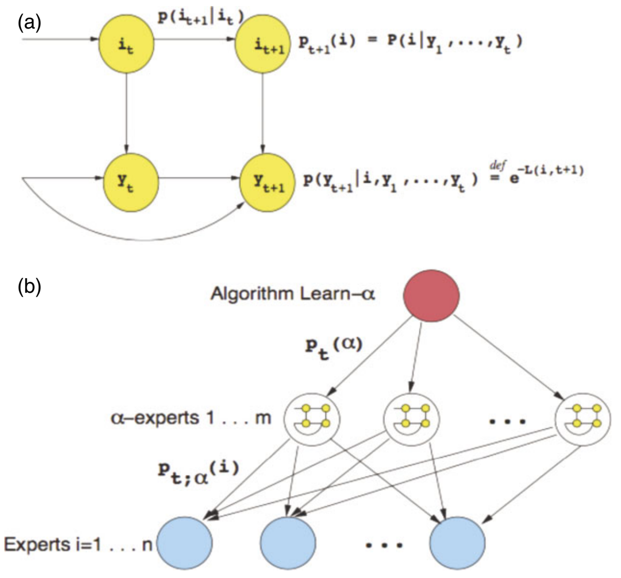

Fig. 2 (a) The generalized Hidden Markov Model corresponding to the algorithms of Herbster and Warmuth [15]. (b) The Learn-α algorithm of Monteleoni and Jaakkola [14]. The α-experts are Fixed-Share(α) algorithms from Herbster and Warmuth [15]. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

transition matrix 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 previous work [14,38] 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 panel (b) of Fig. 2, 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.

3.1. Regret-Optimal Parameter Discretization

We use a discretization procedure for the parameter α given in [14] 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 (as well as different discretizations for the monthly data experiments). Recent work has further studied the issues of discretizing an analogous parameter for similar algorithms [39].

4. GLOBAL EXPERIMENTS

Here we describe the data and experiments at the global scale. In Section 5 we describe our experiments on several geographical regions.

4.1. Global 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 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 [40], and the temperature anomalies are available from NASA [41].) 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 changes 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

```
Algorithm Learn-$\alpha$ 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 $\alpha$ parameter.
Initialization:
  $\forall j,\ p_1(j) \leftarrow \frac{1}{m}$
  $\forall i, j,\ p_{1,j}(i) \leftarrow \frac{1}{n}$
Upon $t$th data observation, $y_t$:
For each $i \in \{1 \ldots n\}$:
  Loss$[i] \leftarrow (y_t - M_i(t))^2$
For each $j \in \{1 \ldots m\}$:
  LossPerAlpha$[j] \leftarrow -\log \sum_{i=1}^{n} p_{t,j}(i)\, e^{-\mathrm{Loss}[i]}$
  $p_{t+1}(j) \leftarrow p_t(j) e^{-\mathrm{LossPerAlpha}[j]}$
  For each $i \in \{1 \ldots n\}$:
    $p_{t+1,j}(i) \leftarrow \sum_{k=1}^{n} p_{t,j}(k)\, e^{-\mathrm{Loss}[k]}\, P(i|k; \alpha_j)$
  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$[j]$
```

Fig. 3 Algorithm Learn-α, due to Monteleoni and Jaakkola [14], applied to tracking climate models.

times in a year, yielding one value for global mean temperature anomaly per year. (In this case the benchmark is averaged over 1951–1980; 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 panel (a) of Fig. 4), 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.

4.2. 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, while others have up to seven different simulation runs (also known as ensemble members). We arbitrarily picked one run per model, for each of the 20 models, as input to all the algorithms. We did so because using all the runs per model would have overemphasized certain models that had substantially more simulation runs. We also 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. The present setting addresses structural uncertainty (among different climate models), rather than initial condition uncertainty (among different simulation runs of one climate model), although the latter topic would be interesting to explore in future work.

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

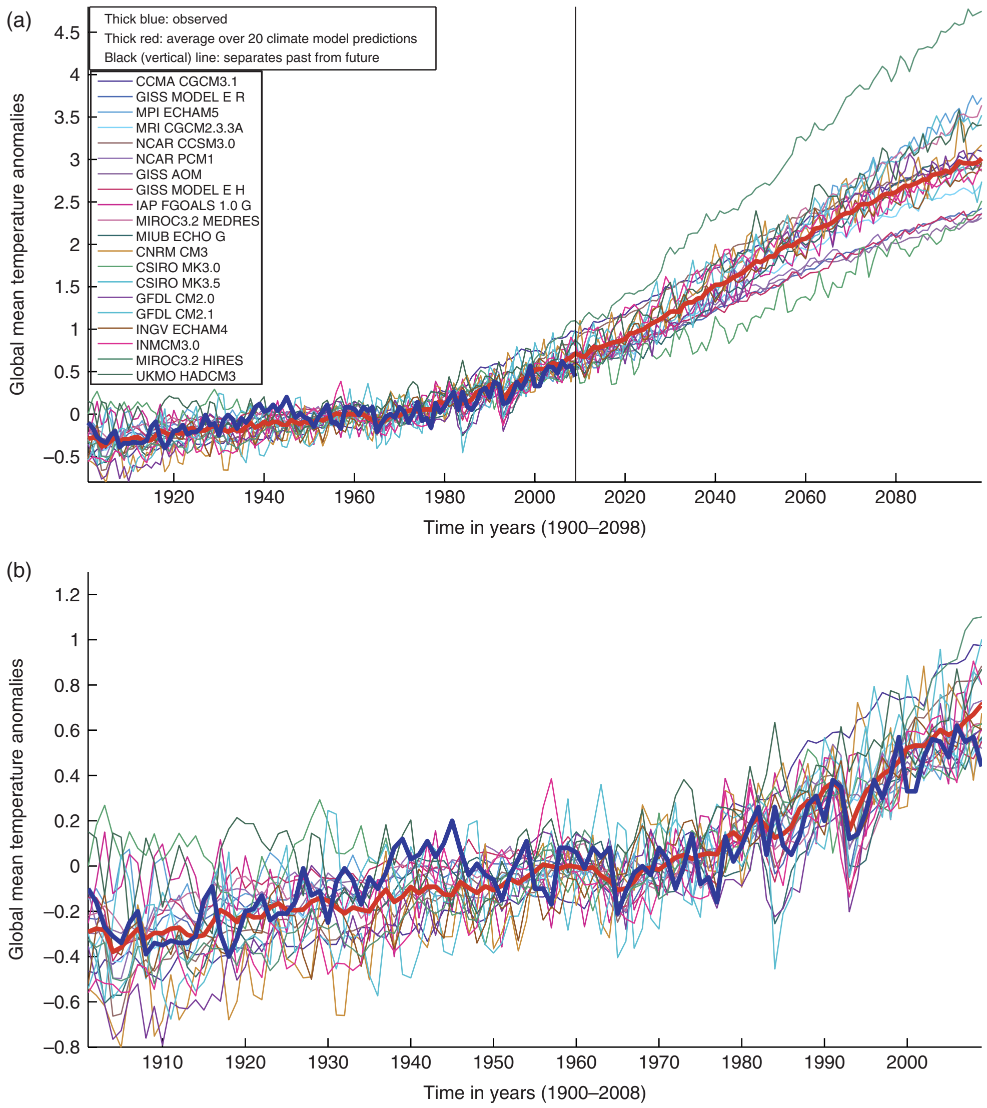

Fig. 4 (a) Model predictions through 2098, with observations through 2008. The black vertical line separates past (hindcasts) from future predictions. (b) Zooming in on observations and model predictions through 2008. The legends refer to both figures. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Institute (Japan), and National Center for Atmospheric Research (Colorado), among others.

4.3. Experiments and Results on Global Data

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 the 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

Table 1. Mean and variance of annual losses.

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

Notes: The best score per experiment is given in bold. The Average Prediction over climate models is the benchmark technique.
[a] 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, starting from the 21st (20th in future simulations) year.

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 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 climate model for each experiment. 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 to a form of test error (in the batch setting). In particular, this "progressive validation" error was analyzed in [42], which provided formal bounds relating it, as well as k-fold cross-validation error, to standard batch holdout error, in certain settings. Thus it is formally related to methods designed to reduce overfitting bias in the evaluation. We also ran sanity-check experiments to verify that Learn-α significantly outperforms the Fixed-Share(α) algorithm, for every value of α in our discretization.

Learn-α's performance, with respect to the average over all model predictions, is very significant, as that is the standard benchmark in climate science. 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. We also notice a general trend that many of the methods do better at predicting in simulations in which the model generating labels performed better historically. In the case of Learn-α, this suggests that the historically poorer models may be relative "outliers", and thus harder to predict using convex combinations of the remaining models' predictions.

Remark. An interesting result is that on global historical data, the best climate model outperforms the average prediction over climate models. While we did not find this to be the case on most of our regional results (and this effect disappears entirely on monthly regional data), the global result appears to contradict the related work in climate science [1,2]. Reichler and Kim [1] 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 [2] 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

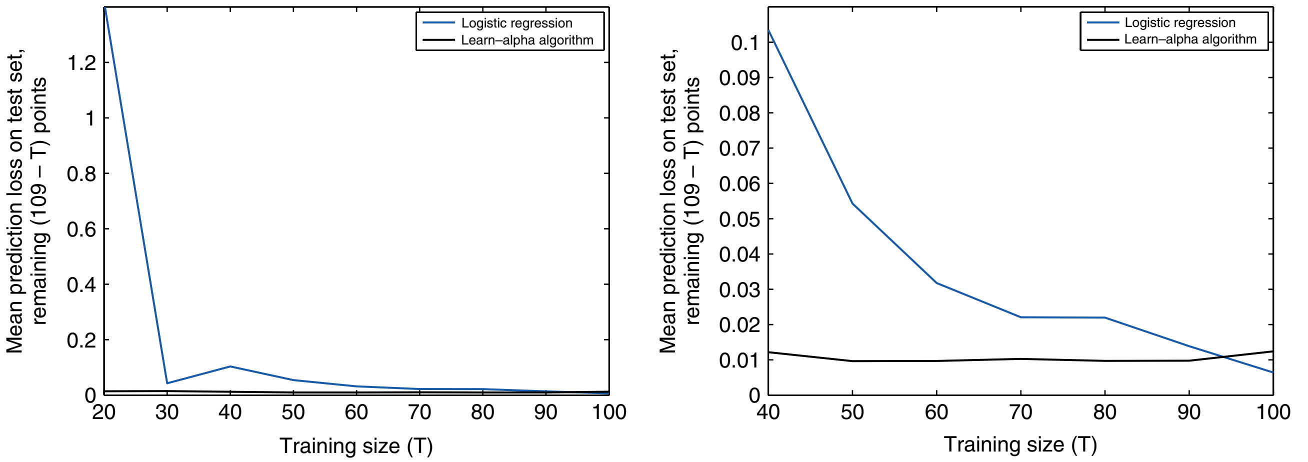

Fig. 5 Batch evaluations. Plot of mean test error on the remaining points, when only the first T are used for training. Right plot zooms in on T $\geq$ 40 (x-axis). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

model that performs better than the average, the "best" expert cannot be used as a 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.

4.4. 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 set 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 ($\sigma^2 = 0.0185$). 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 shown 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, Fig. 5 shows 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.

5. EXPERIMENTS AT HIGHER SPATIAL AND TEMPORAL GRANULARITY

5.1. Regional Data

We also ran experiments at higher spatial and temporal granularity. The global annual data set is generated by averaging the climate models' predictions over the whole globe; here we drilled down on several smaller geographical regions. We used hindcasts of the IPCC global climate models, and the analogous true observations, over specific geographical regions corresponding to several continents, at monthly and annual time-scales. The quantity to predict was still a temperature anomaly; however, the averaging is over a smaller geographical region than the whole globe; in particular, we ran experiments for Africa, Europe, and North America. While the annual experiments have anomaly values averaged over a whole year, we also ran experiments using monthly averages in each of the regions.

The experimental set-up was similar to the global experiments, other than further details explained here. The regions were "boxes" in latitude and longitude corresponding to Europe (0 to 30 E, 40 to 70 N), Africa (-15 to 55 E, -40 to 40 N), and North America (-60 to -180 E, 15 to 70 N). The global climate model projection data (restricted to these regions) was obtained from the KNMI Climate Explorer [43]. The temperature observation data for these regions was attained from NASA [41]. For these queries, we only received climate model data for 19 models, and thus the ensemble size is 19. This data set contains multiple runs per model; we used one run per model, determined randomly. Both model data and observation data are from year 1900 through 2009 (110 years), for the annual experiments and from January 1900 through October 2010 (1330 months) for the historical monthly experiments. We also used monthly regional model predictions through the year 2098, to run future simulations on 2376 months (starting in 1900). For our annual experiments, the preprocessing technique, to ensure that model predictions and observations are both anomalies with respect to the same benchmark period, is analogous to that used in the global experiments, with a benchmark period of 1951–1980. It is important to note that regional data generally has higher variance than the corresponding global data, as each measurement (both predicted and observed) is averaged over fewer geographical regions. For the monthly data, the observations from [41] had already been smoothed to account for seasonality. To match this in the climate model predictions, we preprocessed them using standard techniques of computing anomalies per calendar month. That is, for each monthly measurement, we subtracted the mean over only that particular calendar month (e.g. April), over the benchmark period, 1951–1980. Even after this preprocessing to remove seasonality, the monthly data sets (both predicted and observed) generally has higher variance than annual data for the same region, as each annual measurement is averaged over 12 monthly measurements.

Figure 6 shows the annual temperature anomaly data for each region. The predictions of 19 climate models of the annual mean temperature anomaly over the region in question, are plotted in thin lines, with their average prediction in thick red, and the observed anomalies plotted in thick blue. Figure 7 shows the monthly temperature anomaly data for each region, including into the future, through year 2098. The predictions of 19 climate models are plotted in thin lines, with their average prediction in thick black. Notably, both at annual and monthly time-scales, there are significant differences among the regions.

5.2. Results on Regional Data

In Table 2, we compare mean loss on real-time predictions, i.e. predictions per year, of the algorithms, using the same progressive validation technique as in Table 1. In contrast to the global experiments, there is a general trend for all methods to perform slightly worse and for the variances to increase; indeed as explained in Section 5.1, the data itself generally has higher variance at the regional level than at the global level. The Learn-α algorithm outperforms the Average Prediction of the climate models, which was the state-of-the-art benchmark, as well as linear regression. It also outperforms the best climate model per experiment, except for Africa, in which the performance is close. Notably, the identity of the right climate model for future observations cannot be known in advance. In the annual historical experiments, we found the best climate model per region to be distinct: `giss model e r` for Africa, `miroc3 2 hires` for Europe, and `giss aom` for North America.

In Table 3, we compare mean loss on real-time predictions, predictions per *month* in this case, of the algorithms, using the same progressive validation technique as shown in Table 1. In contrast to the regional annual experiments, there is a general trend for all methods to perform slightly worse and for the variances to increase; as explained in Section 5.1, monthly data generally has higher variance than the corresponding annual data. The Learn-α algorithm is the best performer, notably outperforming both Linear Regression and the Average Prediction over climate models on all experiments, as well as the best climate model per experiment, and the other methods. This result suggests that our average results are robust to scaling up the data set size by more than an order of magnitude; however, the variances increase, as discussed above. In the monthly historical experiments, we found the best climate model per region to be distinct: `miroc3 2 medres` for Africa, `cnrm cm3` for Europe, and `giss model e h` for North America.

We also performed two future simulations per region, using the "perfect model" assumption described in the global experiments. Per region, the simulations took labels from (1) the best model for that historical monthly experiment and (2) the worst model for that historical monthly experiment. The best model identities for the monthly historical experiments are listed above. The second simulation per region used labels from `mcsiro mk3 5` for Africa, `gfdl cm2 0` for Europe, and `giss aom` for North America. The finding that `giss aom` is the worst climate model for the monthly North American experiment, the region for which it was the best climate model for the annual data, is not contradictory; the squared loss is convex, and thus for any climate model, the average loss over a year's worth of its monthly predictions can exceed the loss on its annual prediction, where the annual prediction is averaged over monthly predictions.

In Table 4, we compare mean loss on real-time predictions, predictions per *month* in this case, of the algorithms, using the same progressive validation technique as in Table 1, for the six future simulations, two per region. The Learn-α algorithm's comparative success in average predictions scales to these experiments which have 20× as much data as the annual historical experiments, although the variances again increase, and in one simulation, Linear Regression's performance is better. We also notice

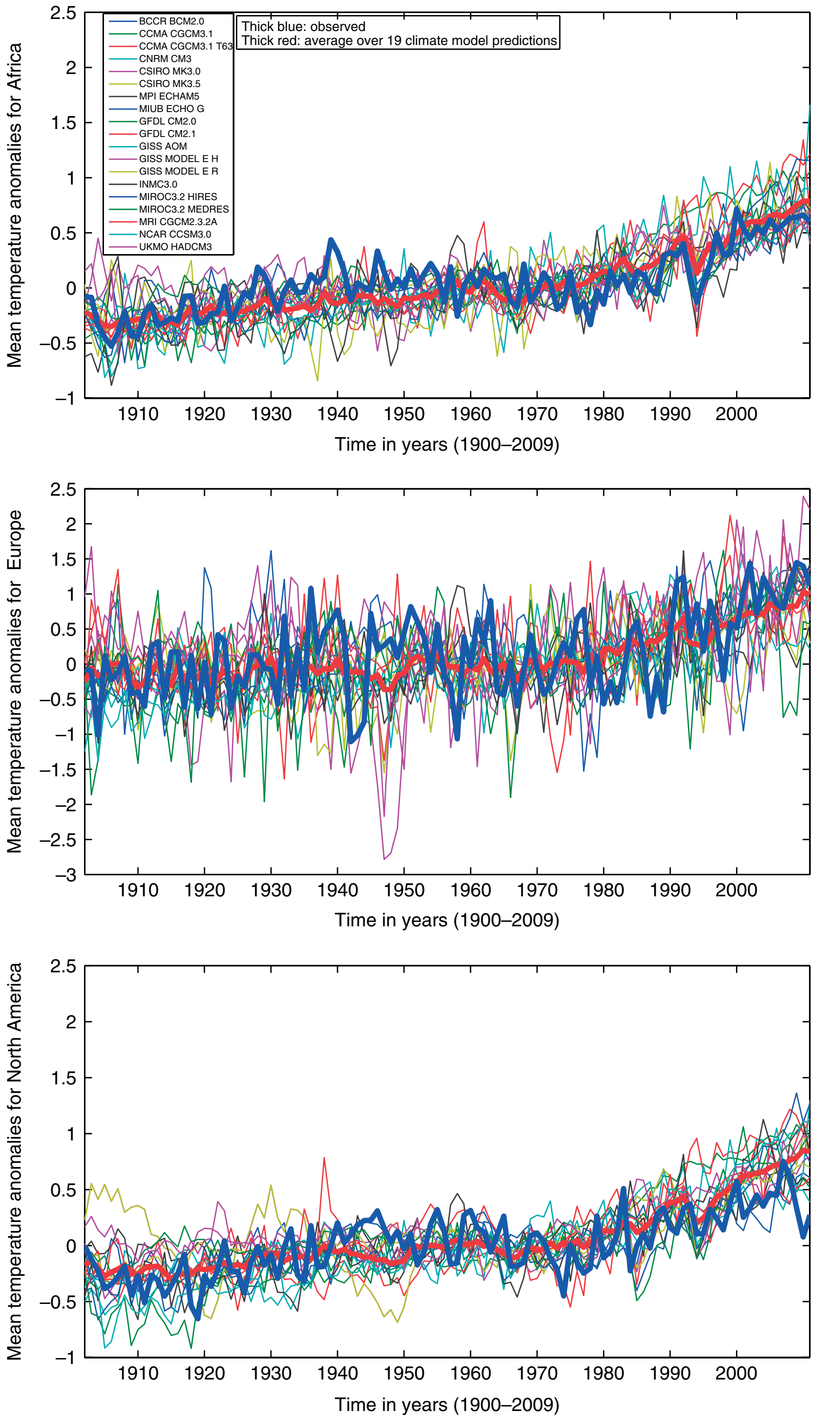

Fig. 6 Annual data sets per region. Legends in top figure apply to all figures. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

a general trend that most methods do better in the first simulation per region, in which the labels are generated by the best historical climate model, than the second, in which the labels are generated by the worst historical climate model. As in the global, annual future simulations, this suggests that the worst predicting climate model, per monthly regional experiment, may vary significantly from the rest of the models, thus making its predictions harder to predict by using convex combinations of the remaining climate models, as Learn-α does.

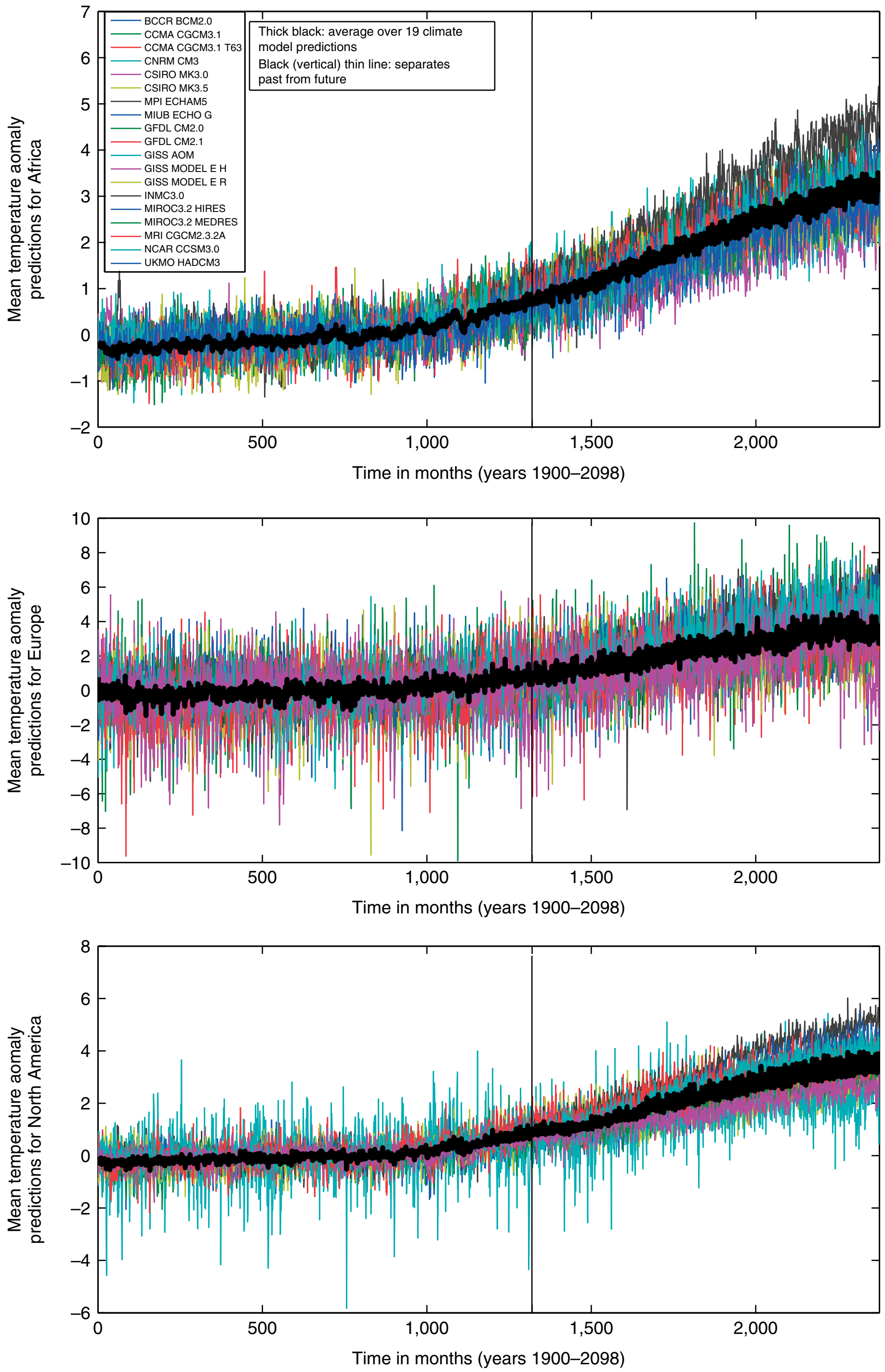

Fig. 7 Monthly model predictions per region, into the future. Legends in top figure apply to all figures. The vertical black line separates the past (hindcasts) from the future predictions. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

6. LEARNING CURVES

Here we provide various learning curves. For the global experiments, we plot the losses of Learn-α against those of 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 8 shows the plot of the squared error between predicted and observed annual mean temperature, by year from 1900 to 2008. Learn-α suffers less loss than

Table 2. Regional results on annual historical data. Mean and variance of annual losses.

Algorithm	Africa	Europe	North America
Learn-α algorithm	0.0283	**0.1794**	**0.0407**
	$\sigma^2 = 0.0020$	$\sigma^2 = 0.0520$	$\sigma^2 = 0.0036$
Linear regression[a]	0.0391	38.9724[b]	0.0704
	$\sigma^2 = 0.0039$	$\sigma^2 = 134700.0$	$\sigma^2 = 0.0156$
Best climate model	**0.0254**	0.2752	0.0450
(for the observations)	$\sigma^2 = 0.0015$	$\sigma^2 = 0.1207$	$\sigma^2 = 0.0035$
Average Prediction	0.0331	0.2383	0.0493
(over climate models)	$\sigma^2 = 0.0025$	$\sigma^2 = 0.0868$	$\sigma^2 = 0.0058$
Median Prediction	0.0291	0.2391	0.0502
(over climate models)	$\sigma^2 = 0.0021$	$\sigma^2 = 0.0964$	$\sigma^2 = 0.0066$
Worst climate model	0.1430	1.0180	0.1593
(for the observations)	$\sigma^2 = 0.0368$	$\sigma^2 = 2.4702$	$\sigma^2 = 0.0372$

Notes: The best score per experiment is given in bold. The Average Prediction over climate models is the benchmark technique.

[a] Linear Regression cannot form predictions for the first 19 years, so its mean is over fewer years than all the other algorithms, starting from the 20th year.

[b] Observing that for Europe, Linear Regression's loss was particularly high on the 20th year, we also computed the mean starting from the 21st year, 0.4989 ($\sigma^2 = 0.4787$). Using this evaluation for Linear Regression, the mean for Africa was 0.0363 ($\sigma^2 = 0.0032$), and the mean for North America was 0.0706 ($\sigma^2 = 0.0158$).

Table 3. Regional results on monthly historical data. Mean and variance of monthly losses.

Algorithm	Africa	Europe	North America
Learn-α algorithm	**0.0598**	**0.3048**	**0.0959**
	$\sigma^2 = 0.0085$	$\sigma^2 = 0.3006$	$\sigma^2 = 0.0311$
Linear regression[a]	0.0741	1.7442	0.1119
	$\sigma^2 = 0.0301$	$\sigma^2 = 43.9616$	$\sigma^2 = 0.0432$
Best expert	0.1144	2.2498	0.1629
(for the observations)	$\sigma^2 = 0.0285$	$\sigma^2 = 15.4041$	$\sigma^2 = 0.0935$
Average Prediction	0.0752	1.4781	0.1101
(over climate models)	$\sigma^2 = 0.0106$	$\sigma^2 = 7.5964$	$\sigma^2 = 0.0417$
Median Prediction	0.0777	1.5001	0.1116
	$\sigma^2 = 0.0117$	$\sigma^2 = 8.1498$	$\sigma^2 = 0.0456$
Worst expert	0.2333	4.2104	1.1698
(for the observations)	$\sigma^2 = 0.1020$	$\sigma^2 = 71.2737$	$\sigma^2 = 6.3192$

Notes: The best score per experiment is given in bold. The Average Prediction over climate models is the benchmark technique.

[a] Linear regression cannot form predictions for the first 19 months, so its mean is over fewer months than all the other algorithms, starting from the 20th month.

the mean over model predictions on over 75% of the years (82/109).

The learning curves for the global future simulation experiments (Figs 9 and 10) 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 global 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.

We also provide learning curves for the regional future simulations which generated the statistics summarized in Table 4. These simulations also started at 1900, but since the data is at a monthly time-scale, the figures zoom in on the period from 2009 to 2098. Figure 11 compares the monthly losses of the Learn-α algorithm (black), to those of the average prediction over the climate models (red), for the first simulation per region, in which the best historical climate model provides the labels. Figure 12 does so for the second simulation per region, in which the worst historical climate model provides the labels. In keeping with the results shown in Table 4, it is apparent that the online learning algorithm suffers less prediction loss in each experiment than the benchmark method, the average prediction over climate models.

Table 4. Regional results on two future simulations per region. Mean and variance of monthly losses.

Algorithm	Africa 1	Africa 2	Europe 1	Europe 2	N. Amer. 1	N. Amer. 2
Learn-α algorithm	**0.0890** $\sigma^2 = 0.0167$	**0.1053** $\sigma^2 = 0.0249$	**0.2812** $\sigma^2 = 0.4134$	**0.6624** $\sigma^2 = 3.6678$	0.0968 $\sigma^2 = 0.0272$	**0.6061** $\sigma^2 = 1.6429$
Linear regression[a]	0.0985 $\sigma^2 = 0.2680$	0.1384 $\sigma^2 = 0.0455$	1.1487 $\sigma^2 = 4.2672$	3.0836 $\sigma^2 = 44.1931$	**0.0923** $\sigma^2 = 0.0365$	1.0458 $\sigma^2 = 4.4447$
Best expert (for the observations)	0.1912 $\sigma^2 = 0.0757$	0.1967 $\sigma^2 = 0.0754$	2.1210 $\sigma^2 = 12.6767$	3.7893 $\sigma^2 = 39.2087$	0.1713 $\sigma^2 = 0.0903$	1.0478 $\sigma^2 = 3.9090$
Average Prediction (over climate models)	0.1388 $\sigma^2 = 0.0410$	0.1806 $\sigma^2 = 0.0716$	1.1106 $\sigma^2 = 4.4023$	2.9353 $\sigma^2 = 29.9128$	0.1432 $\sigma^2 = 0.0478$	1.0745 $\sigma^2 = 4.1346$
Median Prediction	0.1266 $\sigma^2 = 0.0352$	0.1711 $\sigma^2 = 0.0637$	1.1385 $\sigma^2 = 4.5734$	2.9093 $\sigma^2 = 30.3332$	0.1835 $\sigma^2 = 0.0827$	1.1075 $\sigma^2 = 4.2544$
Worst expert (for the observations)	0.5236 $\sigma^2 = 0.5782$	0.5625 $\sigma^2 = 0.7018$	3.8266 $\sigma^2 = 47.7359$	5.0029 $\sigma^2 = 76.7785$	1.2311 $\sigma^2 = 3.3160$	2.2641 $\sigma^2 = 12.0301$

Notes: The best score per experiment is given in bold. The Average Prediction over climate models is the benchmark technique.
[a]Linear Regression cannot form predictions for the first 18 months, so its mean is over fewer months than all the other algorithms, starting from the 19th month.

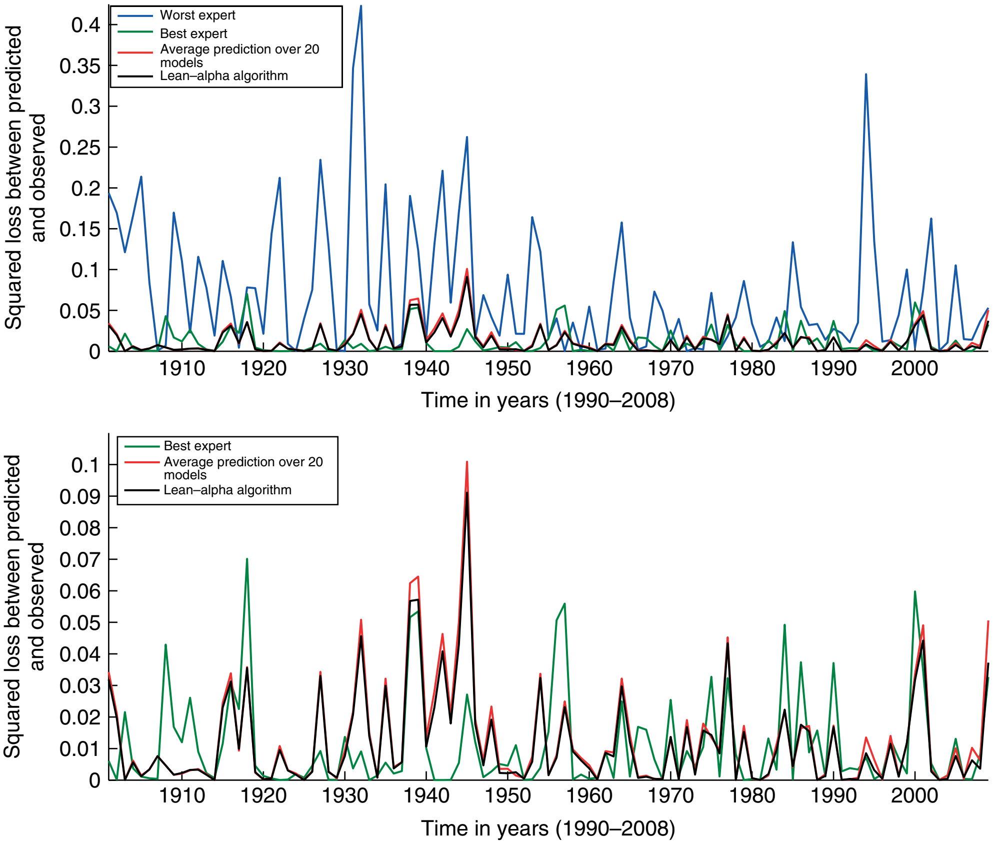

Fig. 8 Historical Global Annual experiment. Squared loss between predicted and observed global mean temperature anomalies. The bottom plot zooms in on the y-axis. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

6.1. 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 global historical data experiment.

Panel (a) of Fig. 13 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 α consistently

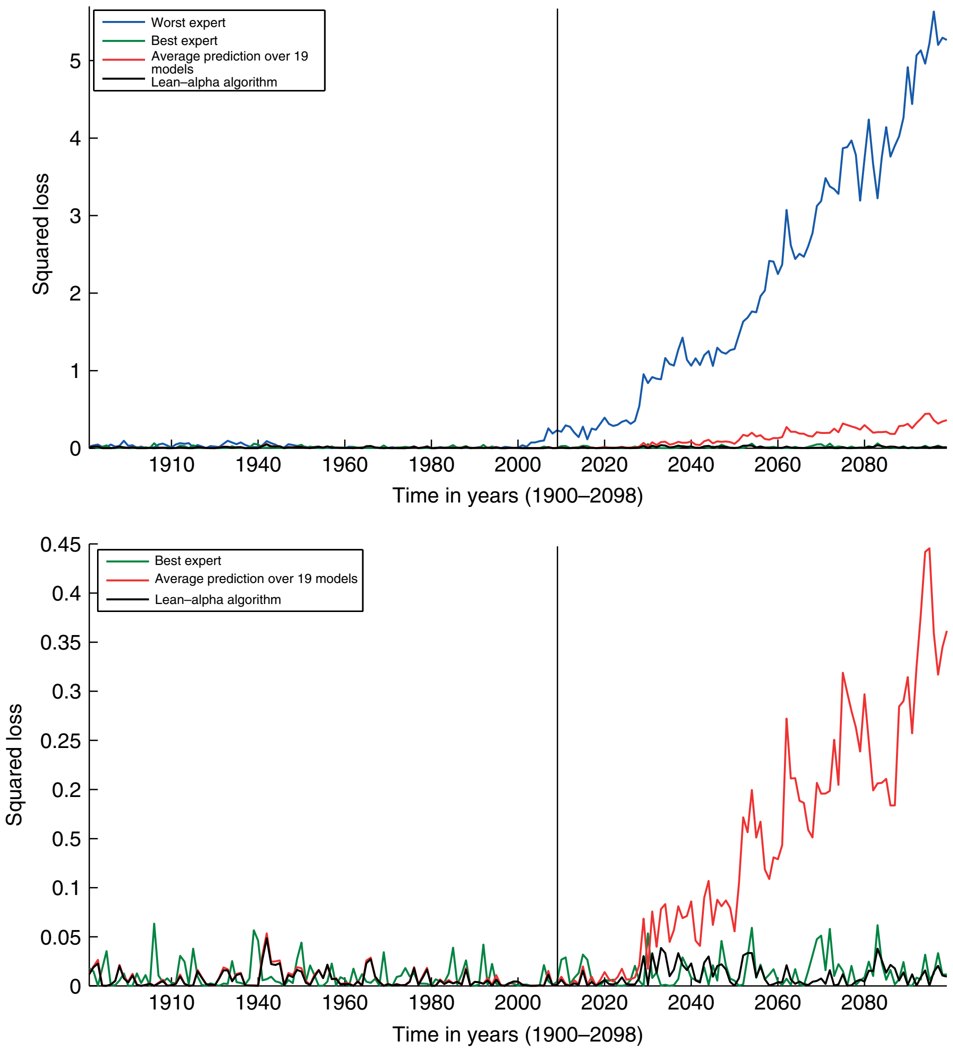

Fig. 9 Global Future Simulation 1: tracking the predictions of one model using the predictions of the remaining 19 as input, with no true temperature observations. Black vertical line separates past (hindcasts) from future predictions. The bottom plot zooms in on the y-axis. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

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 annual historical data experiments.

Panel (b) of Fig. 13 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 global historical data set.

7. DISCUSSION AND FUTURE WORK

These encouraging results will hopefully lead to a fuller exploration of whether there is enough information in comparison of model hindcasts to observations to assess projection credibility. Climate model projections have inherent uncertainties related to internal chaotic variability, structural uncertainty related to our incomplete understanding of the climate system, and scenario uncertainty related to the impossibility of knowing exactly how economies, technologies, and regulatory frameworks that impact emissions will change in the future [19]. Comparisons of past climate to model hindcasts predominantly provide information related to structural uncertainties, though some predictability related to the internal variability may also be derivable; this is an interesting direction for future work.

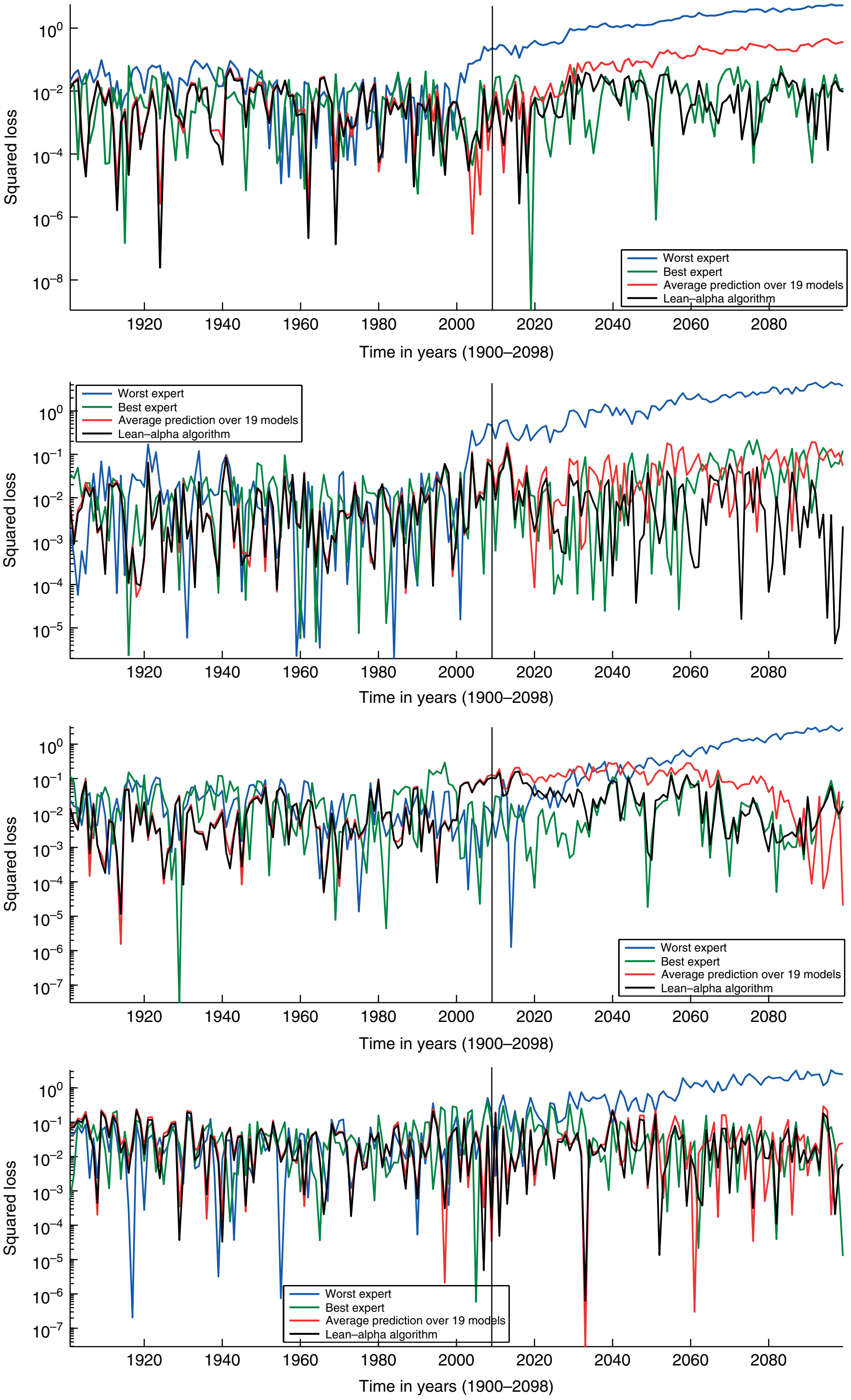

Fig. 10 Global Future Simulations. From top to bottom: 1, 2, 3, and 4, in log-scale. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

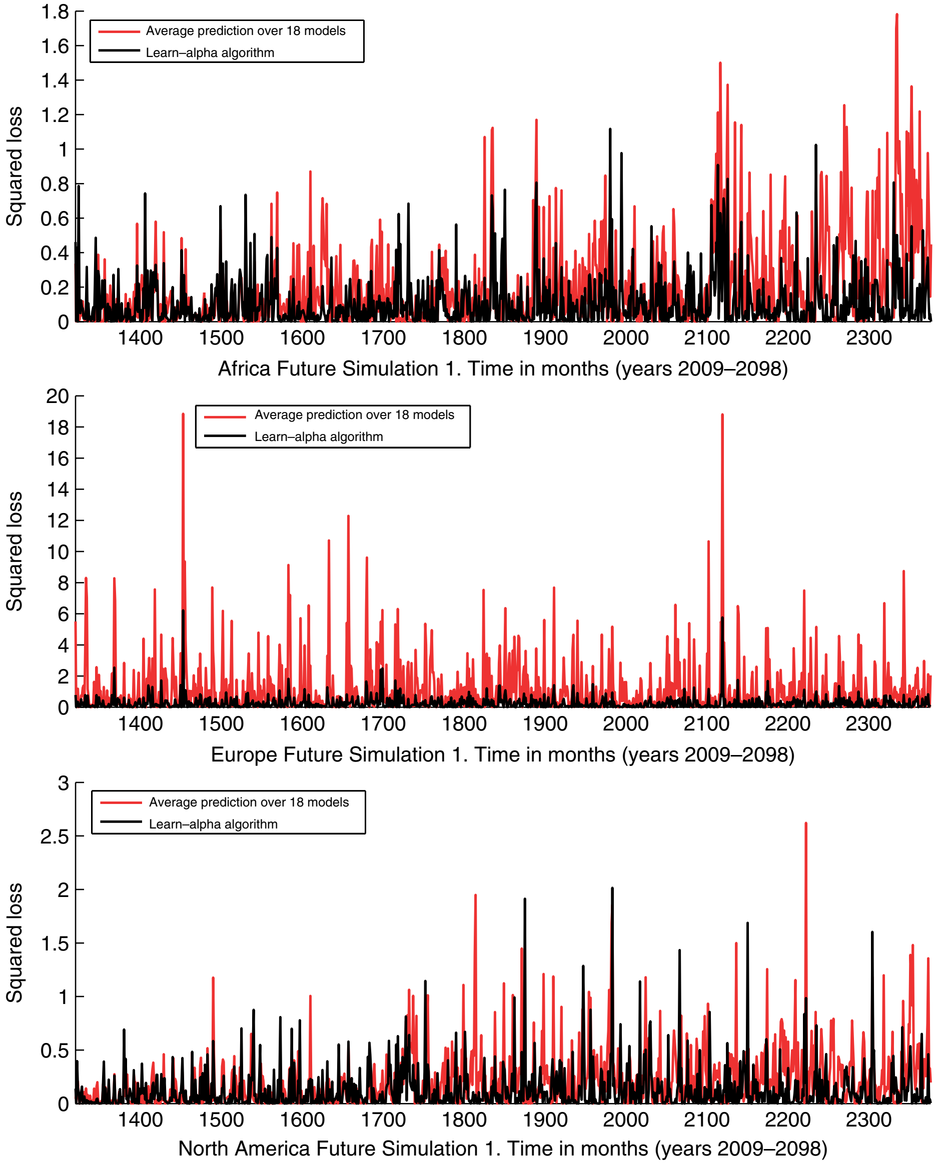

Fig. 11 Future Regional Monthly Simulations (1). The simulations start at 1900, but the plots start at 2009 to zoom in on the future. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

An important next challenge, is how to track climate models when predicting *future* climates. Existing tracking methods rely on receiving true observations, with which to evaluate the models' predictions. One goal for future work in the design of machine learning and data mining algorithms, would be to 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 [44]. There is some literature on regret analyses of semi-supervised online learning; Cesa-Bianchi *et al*. [45,46] 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. [47]. One approach that we have shown to be feasible in practice (in our future simulations) is to view expert predictions themselves as partial feedback, in order to design semi-supervised algorithms. We can also turn to the batch setting, in which future predictions are needed, given all past data, and exploit other areas of the machine learning and data mining literature.

In summary, our results advance the state-of-the-art in the climate science community, with respect to combining climate model predictions. Our approach in this work

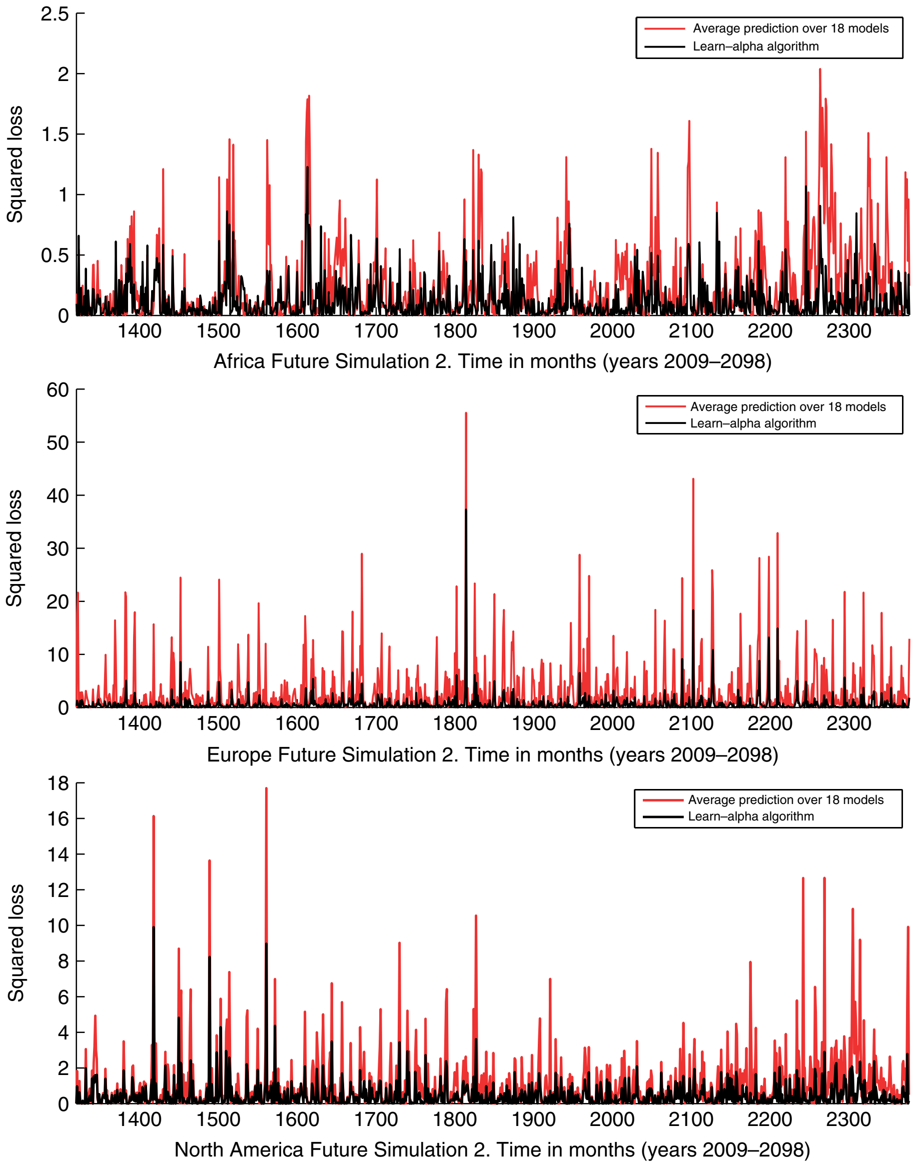

Fig. 12 Future Regional Monthly Simulations (2). The simulations start at 1900, but the plots start at 2009 to zoom in on the future. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

has significant qualitative differences from most current practices in climate science; the ensemble weights are updated adaptively, and the frequency of switches among which climate model predicts best at any given time, is also learned simultaneously. We have shown the applicability of our techniques both at global and regional time-scales, and for annual and monthly predictions. Our methods are applicable to any quantity predicted by a set of climate models, and we plan to use them for predicting other important climate benchmarks, such as concentrations of carbon dioxide and other greenhouse gases. In future work, it would also be interesting to look at smaller time-scales, using different aggregation techniques, as well as to consider other geographical regions. There remains a rich source of unexplored information in the paleo-climate record and in multiple other data sets over the instrumental period that could be used to track climate model performance and provide more informative projections. The work of assessing which metrics, or combinations thereof, provide the most information on model predictions has barely begun, and this is an interesting area for future research. In addition to our specific contributions, we hope to inspire future applications of machine learning and data mining to improve climate predictions and to help answer pressing questions in climate science.

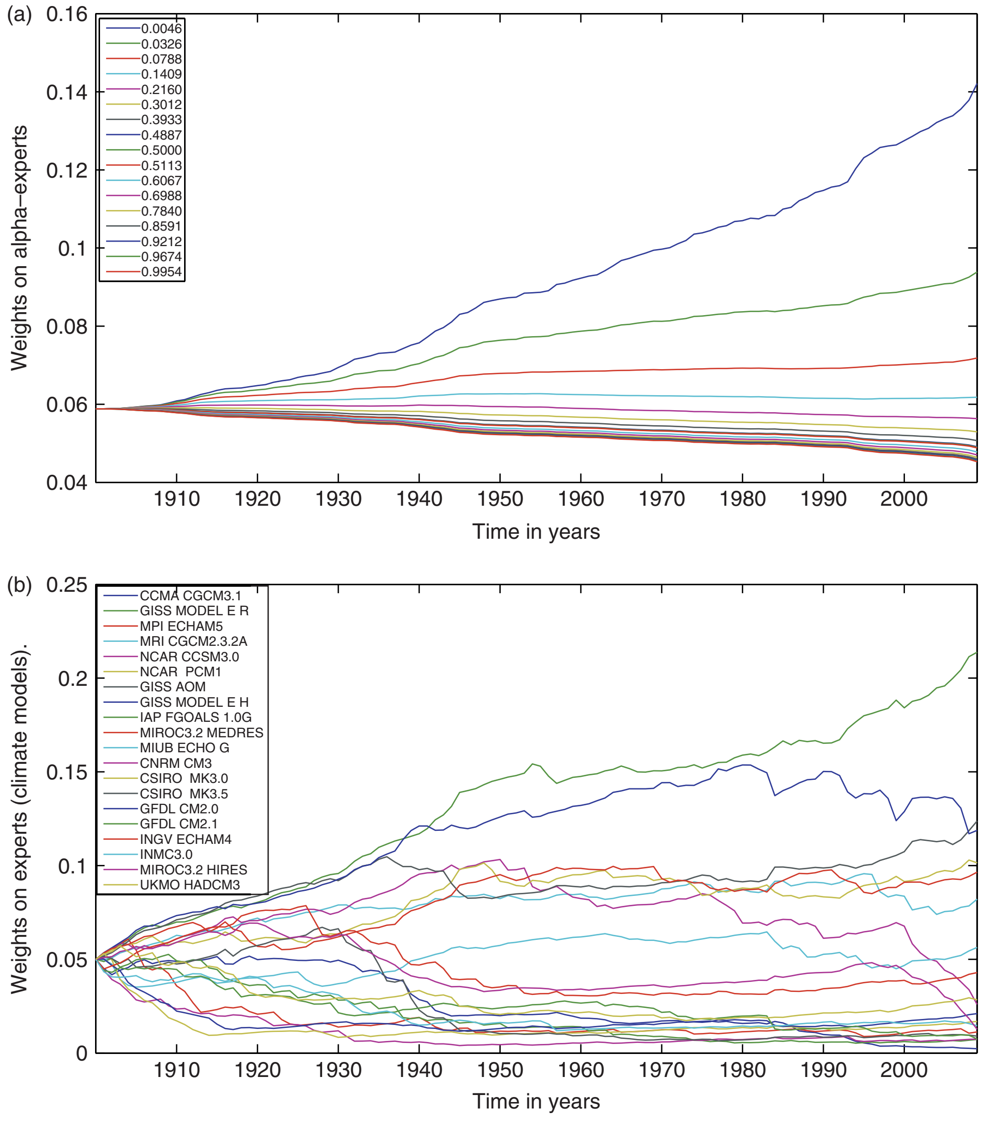

Fig. 13 Weight evolution. (a) Algorithm's weights on α-experts. Legend lists α-values. (b) Best α-expert's weights on experts (climate models). Legend lists climate models. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

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 their helpful comments on earlier versions of this work. We also thank the anonymous journal reviewers for their comments used in revision.

We would like to acknowledge funding from the Earth Institute at Columbia University, which supported the work of E.A. We also acknowledge the climate modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI), and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, US Department of Energy.

REFERENCES

[1] T. Reichler and J. Kim, How well do coupled models simulate today's climate? Bull Am Met Soc 89 (2008), 303–311.

[2] C. Reifen and R. Toumi, Climate projections: past performance no guarantee of future skill? Geophys Res Lett 36 (2009).

[3] 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, Int J Intell Syst 20(7) (2005), 759–787.

[4] 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, 2007, 252–261.

[5] V. Kumar, Discovery of patterns in global earth science data using data mining. In PAKDD (1), 2010.

[6] A. Braverman, R. Pincus, and C. Batstone, Data mining for climate model improvement, In 6th Annual NASA Earth Science Technology Conference, 2006.

[7] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter, Discovery of climate indices using clustering, In KDD '03: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, 446–455.

[8] C. Lima, U. Lall, T. Jebara, and A. G. Barnston, Statistical prediction of ENSO from subsurface sea temperature using a nonlinear dimensionality reduction, J Clim 22(17) (2009), 4501–4519.

[9] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly, An exploration of climate data using complex networks, ACM SIGKDD Explor 12(1) (2010), 25–32.

[10] 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, 2009, 587–596.

[11] 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 Netw 19(2) (2006), 122–134.

[12] V. M. Krasnopolsky, M. S. Fox-Rabinovitz, and A. A. Belochitski, Decadal climate simulations using accurate and fast neural network emulation of full, longwave and shortwave, radiation, Mon Weather Rev 136 (2008), 368–3695.

[13] P. J. Roebber, M. R. Butt, S. J. Reinke, and T. J. Grafenauer, Real-time forecasting of snowfall using a neural network, Weather Forecast 22(3) (2007), 676–684.

[14] C. Monteleoni and T. Jaakkola, Online learning of non-stationary sequences, In NIPS '03: Advances in Neural Information Processing Systems 16, 2003.

[15] M. Herbster and M. K. Warmuth, Tracking the best expert, Mach Learn 32 (1998), 151–178.

[16] http://celebrating200years.noaa.gov/breakthroughs/climate_model/welcome.html.

[17] G. A. Schmidt, R. Ruedy, J. E. Hansen, I. Aleinov, N. Bell, M. Bauer, S. Bauer, B. Cairns, V. Canuto, Y. Cheng, A. Del Genio, G. Faluvegi, A. D. Friend, T. M. Hall, Y. Hu, M. Kelley, N. Y. Kiang, D. Koch, A. A. Lacis, J. Lerner, K. K. Lo, R. L. Miller, L. Nazarenko, V. Oinas, J. Perlwitz, J. Perlwitz, D. Rind, A. Romanou, G. L. Russell, M. Sato, D. T. Shindell, P. H. 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, J Clim 19 (2006), 153–192.

[18] 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 Natl Acad Sci U S A 106 (2009), 14778–14783.

[19] E. Hawkins and R. Sutton, The potential to narrow uncertainty in regional climate predictions, Bull Am Met Soc 90 (2009), 1095–1107.

[20] S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, eds. 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, New York, Cambridge University Press, 2007.

[21] R. Knutti, The end of model democracy? Clim Change 102 (2010), 395–404.

[22] R. Knutti, R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, Challenges in combining projections from multiple climate models, J Clim 23(10) (2010), 2739–2758.

[23] 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.

[24] J. D. Annan and J. C. Hargreaves, Reliability of the CMIP3 ensemble, Geophys Res Lett 37 (2010), L02703.

[25] J. D. Annan and J. C. Hargreaves, Understanding the CMIP3 multi-model ensemble, J Climate (2011), in press

[26] D. A. Stainforth, M. R. Allen, E. R. Tredger, and L. A. Smith, Confidence, uncertainty and decision-support relevance in climate predictions, Philos Trans R Soc A: Math Phys Eng Sci 365(1857) (2007), 2145–2161.

[27] D. Klocke, R. Pincus, and J. Quaas, On constraining estimates of climate sensitivity with present-day observations through model weighting, J Clim (2011), in press.

[28] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski, Using Bayesian model averaging to calibrate forecast ensembles, Mon Weather Rev 133 (2005), 1155–1174.

[29] A.M. Greene, L. Goddard, and U. Lall, Probabilistic multimodel regional temperature change projections, J Clim 19(17) (2006), 4326–4343.

[30] T. DelSole, A Bayesian framework for multimodel regression, J Clim 20 (2007), 2810–2826.

[31] M. K. Tippett and A. G. Barnston, Skill of multimodel ENSO probability forecasts, Mon Weather Rev 136 (2008), 3933–3946.

[32] M. Peña and H. van den Dool, Consolidation of multimodel forecasts by ridge regression: Application to pacific sea surface temperature, J Clim 21 (2008), 6521–6538.

[33] S. Casanova and B. Ahrens, On the weighting of multimodel ensembles in seasonal and short-range weather forecasting, Mon Weather Rev 137 (2009), 3811–3822.

[34] R. L. Smith, C. Tebaldi, D. Nychka, and L. O. Mearns, Bayesian modeling of uncertainty in ensembles of climate models, J Am Stat Assoc 104(485) (2009), 97–116.

[35] S. Sain and R. Furrer, Combining climate model output via model correlations, Stochastic Environ Res Risk Assess (2010).

[36] N. Littlestone and M. K. Warmuth, The weighted majority algorithm, In Proceedings IEEE Symposium on Foundations of Computer Science, 1989, 256–261.

[37] N. Cesa-Bianchi and G. Lugosi, Prediction, learning, and games, Cambridge University Press, 2006.

[38] C. E. Monteleoni, Online learning of non-stationary sequences, SM Thesis. MIT Artificial Intelligence Technical Report 2003-011, 2003.

[39] 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.

[40] http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php.

[41] http://data.giss.nasa.gov/gistemp/.

[42] 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, 1999, 203–208.

[43] http://climexp.knmi.nl.

[44] O. Chapelle, B. Schölkopf, and A. Zien, eds. Semi-Supervised Learning, Cambridge, MA, MIT Press, 2006.

[45] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz, Minimizing regret with label efficient prediction, IEEE Trans Inform Theory 51(6) (2005), 2152–2162.

[46] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni, Worst-case analysis of selective sampling for linear-threshold algorithms, J Mach Learn Res 7 (2006), 1205–1230.

[47] G. Lugosi, S. Mannor, and G. Stoltz, Strategies for prediction under imperfect monitoring, In Proceedings 20th Annual Conference on Learning Theory, 2007.