THESIS

APPLICATION OF AN INTERPRETABLE PROTOTYPICAL-PART NETWORK TO SUBSEASONAL-TO-SEASONAL CLIMATE PREDICTION OVER NORTH AMERICA

Submitted by

Nicolas J. Gordillo

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2024

Master's Committee:

Advisor: Elizabeth Barnes

Russ Schumacher
Chuck Anderson

ABSTRACT

APPLICATION OF AN INTERPRETABLE PROTOTYPICAL-PART NETWORK TO SUBSEASONAL-TO-SEASONAL CLIMATE PREDICTION OVER NORTH AMERICA

In recent years, the use of neural networks for weather and climate prediction has greatly increased. In order to explain the decision-making process of machine learning "black-box" models, most research has focused on the use of machine learning explainability methods (XAI). These methods attempt to explain the decision-making process of the black box networks after they have been trained. An alternative approach is to build neural network architectures that are inherently interpretable. That is, construct networks that can be understood by a human throughout the entire decision-making process, rather than explained post-hoc. Here, we apply such a neural network architecture, named ProtoLNet, in a subseasonal-to-seasonal climate prediction setting. ProtoLNet identifies predictive patterns in the training data that can be used as prototypes to classify the input, while also accounting for the absolute location of the prototype in the input field. In our application, we use data from the Community Earth System Model version 2 (CESM2) pre-industrial long control simulation and train ProtoLNet to identify prototypes in precipitation anomalies over the Indian and North Pacific Oceans to forecast 2-meter temperature anomalies across the western coast of North America on subseasonal-to-seasonal timescales. These identified CESM2 prototypes are then projected onto fifth-generation ECMWF Reanalysis (ERA5) data to predict temperature anomalies in the observations several weeks ahead. We compare the performance of ProtoLNet between using CESM2 and ERA5 data. We then demonstrate a novel approach for performing transfer learning between CESM2 and ERA5 data which allows us to identify skillful prototypes in the observations. We show that the predictions by ProtoLNet using both datasets have skill while also being interpretable, sensible, and useful for drawing conclusions about what the model has learned.

TABLE OF CONTENTS

Chapter 1

Introduction

The usage of machine learning within the geosciences has continued to steadily increase with each passing year. Neural networks, in particular, have seen intense use as they can discover non-linear relationships from within highly complex data (e.g, Mayer and Barnes 2021; Gordon et al. 2021; Connolly et al. 2023). However, a limitation of most neural networks is that they are "black-boxes" (Rudin et al. 2022). That is, the process of how a network made its predictions is often not comprehensible to a human. This is an issue as many black-box networks are actively applied to make high-stake predictions (Rudin 2019; McGovern et al. 2022). In-turn, this prevents many neural networks from being used by operational agencies, who require a thorough description of the reasoning for any decision from start to finish. One solution to understanding black-box models is to deploy machine learning explainability methods (XAI) on network predictions *post-hoc* (e.g Toms et al. 2020; Molina et al. 2023). That is, while XAI allows for educated and useful inferences regarding a network's decision-making, XAI methods are inherently imperfect as they do not replicate the exact decision-making process/computations done by a complex, deep neural network (Rudin et al. 2022). XAI methods must simplify the decisions by the network which can result in varying explanations by multiple methods and potentially unreliable explanations (Mamalakis et al. 2022). Thus, XAI methods often require significant analysis to ensure sensibility.

An alternative solution is to instead build networks that are inherently interpretable. One example of an interpretable network is ProtoPNet (Chen et al. 2018). The interpretability of ProtoPNet arises from the use of prototypes. ProtoPNet ingests training images and finds patches of prototypical patterns directly from the training images to use as prototypes to then classify testing images.

Thus, it is possible to directly understand the decision-making process by ProtoPNet by examining the sample and prototypes the network used to make the prediction. For example, Chen et al. (2018) used ProtoPNet to classify birds. In doing this, ProtoPNet found distinctive prototypical features that distinguish birds from one another. For example, ProtoPNet identified wing color as a prototypical feature, which was used as a prototype to classify other birds by directly comparing bird wings. ProtoPNet, when classifying birds, had comparable skill to black-box networks trained on the same problem with the added benefit of providing interpretability.

ProtoPNet is location invariant when recognizing similarities between prototype and sample due to its use of convolutional filters without a following fully connected dense layer. That is, it does not matter where the prototype and sample features exist in the input space when making a prediction. While location invariance is useful for many image prediction problems, it can be detrimental in a climate and weather prediction setting. For example, precipitation over the Indian Ocean will have different impacts and eventual evolution than precipitation over the eastern Pacific Ocean (Stan et al. 2017), but the spatial invariance of ProtoPNet will be unable to distinguish a difference. To address this, Barnes et al. (2022) modified ProtoPNet by adding an additional location scaling grid associated with each prototype which they termed ProtoLNet. In doing so, ProtoLNet is able to learn that learned prototypical patterns are only important for certain locations within the input. Thus, a precipitation anomaly over the eastern Pacific Ocean can be explicitly distinguished by ProtoLNet from the same precipitation anomaly over the Indian Ocean. Barnes et al. (2022) applied ProtoLNet to simplified toy problems and demonstrated that it was able to accurately distinguish prototypes by location.

Whereas Barnes et al. (2022) showed that ProtoLNet is successful when applied to idealized identification tasks, it is unknown whether it is successful in a more realistic geophysical setting.

Here, we apply ProtoLNet in a subseasonal-to-seasonal (S2S) climate prediction setting. S2S prediction timescales, typically spanning 2 weeks to a full meteorological season (3 months), have typically been associated with a severe lack of prediction skill (e.g. Vitart et al. 2017). Most weather models have very low skill when forecasting at the S2S timescale (e.g. White et al. 2017), but improvements have been made in recent years (e.g. Son et al. 2020). The tropics are well acknowledged to provide S2S prediction skill through the Madden-Julian Oscillation (MJO; Madden and Julian 1971, 1972), active over the tropical Indian and Pacific Oceans, and El-Niño Southern Oscillation (ENSO; Trenberth 1997), situated near the tropical eastern Pacific Ocean (Stan et al. 2017; Arcodia et al. 2020). Thus, we elect to apply ProtoLNet for S2S prediction in order to leverage ProtoLNet's ability to detect large-scale tropical variability patterns, while also being interpretable.

The prototypical networks developed by Chen et al. (2018) and Barnes et al. (2022) show that while such networks are promising, the prediction problem must be chosen carefully as the network is constrained by using a small subset of prototypical patterns to perform skillful predictions. Our S2S tasks are appropriate as there is ample evidence that skillful predictions come about from only a subset of large-scale climate variability patterns during "forecasts of opportunity" (e.g. Mariotti et al. 2020). Forecasts of opportunity are specific climate states that lend themselves to enhanced prediction skill. Previous studies have shown that forecasts of opportunity can be identified by black box neural networks using the MJO and ENSO teleconnections as the sources of predictability (e.g. Mayer and Barnes 2022; Arcodia et al. 2023). Thus, the allocation of a small subset of prototypical patterns by ProtoLNet lends itself to the forecast of opportunity paradigm. However, none of these previous studies used interpretable models, and thus, were required to rely on XAI methods. In fact, there are very few interpretable networks developed for climate or

weather prediction. Hilburn (2023) found that building an interpretable model required extensive time and effort, but it yielded results that otherwise would have been missed by XAI methods. Thus, our work aims to continue to fill this gap by applying an interpretable network to an S2S climate prediction application.

One source of difficulty with S2S prediction is the lack of observational samples. As deep learning requires large amounts of data, many exploratory S2S prediction studies rely on climate models. However, climate models have their own biases which can make the application of networks trained on climate model data alone unfit for observational use. Transfer learning is one possible solution to this, but it is typically applied to black-box architectures and has only been successful in a few climate prediction cases (e.g. Ham et al. 2019). Since deep learning on large quantities of climate data can also lead to "overfitting" on a climate model's biases, using a network like ProtoLNet may be helpful as ProtoLNet constrains the network by forcing the use of a small subset of prototypical patterns to learn the most important patterns. We explore this more here.

Chapter 2 explains the data used for this study while Chapter 3 outlines the ProtoLNet architecture along with two additional modifications to the original architecture used for observational prediction. Chapter 4 shows how insightful prototypical networks can be through using physical patterns to interpret results from climate model data and observations. A brief discussion and conclusions of this study are found in Chapter 5. This thesis has been written in the style of a journal article and will be submitted to Artificial Intelligence for Earth Systems (AIES), a journal of the American Meteorological Society.

Chapter 2

Data

2.1 CESM2 Data

We use 550 years of daily fields of precipitation (mm/day) and 2-meter temperature (K) from the pre-industrial Community Earth System Model 2 (CESM2) long control simulation (Danabasoglu et al. 2020). These fields are converted into anomalies by subtracting the third-order polynomial trend across all 550 years at each grid point for each calendar day separately (thus removing the seasonal cycle as well as any model drift). To reduce the impact of high frequency "noise", a 5-day backward moving average is applied to the precipitation anomalies (predictor) while a 5-day forward moving average is applied to the temperature anomalies (predictand). We use a pre-industrial long control simulation spanning hundreds of years to ensure ample training, validation, and testing data. Furthermore, by focusing on a long control simulation, we are able to evaluate the performance of our ProtoLNet approach in a stationary climate. The real world, of course, is not stationary. However, we will show that the ProtoLNet trained on the long control simulation (denoted as CESM2-ProtoLNet) exhibits skill when applied to the observations. Future work could explore whether using historical simulations improves the overall performance of ProtoLNet for our S2S prediction task.

As both the MJO and ENSO are important sources of S2S predictability across North America (Huang et al. 2021), we focus on precipitation fields between 30°N to 30°S and 40°E to 60°W for our predictors. Since MJO teleconnections are strongest during the boreal winter, we use precipitation anomalies from November 1st through February 28th as input to ProtoLNet. Similarly,

for our predictand, we examine 2-meter temperatures across 18 locations on the western coast of North America where MJO and ENSO teleconnections from the tropics are well documented (Figure A.1). We predict temperatures at a 14-day lead time (14-18 days with the 5-day moving average window) and these dates therefore span the adjusted boreal winter season of 15 November to 14 March. Temperature anomalies at each location are classified into one of three classes. The cold class is defined as temperatures below or equal to the 33rd percentile of all temperature anomalies for that location. Likewise, the warm class is defined as all temperatures above the 66th percentile. The neutral class is defined as all temperatures between the 33rd and 66th percentiles. The data is shuffled by boreal winter season, utilizing the same random seed used to train the base CNN and ProtoLNet (see Chapter 3), and split into 350 years for training, 100 years for validation, and 100 years for testing.

2.2 ERA5 Data

Although ProtoLNet trained on CESM2 data (CESM2-ProtoLNet) is a large focus of this work, a major result of this study is demonstrating how CESM2-ProtoLNet can be applied to observations. We use the European Centre for Medium-Range Weather Forecasts Reanalysis version 5 (ERA5) data as our observational dataset (Hersbach et al. 2020). ERA5 data allows for ProtoLNet to identify and use prototypical features that have occurred on specific dates in the real world. We use 62 years of ERA5 data starting in 1959 and ending in 2021. Although ERA5 now extends back to 1940, daily values were not readily available at the time of this analysis. CESM2-ProtoLNet requires the grid resolution of any new input to match that of the processed CESM2 input, $\approx .94^\circ$ latitude by 1.25° longitude. Therefore, we utilized the Max-Planck-Institute for Meteorology's re-

gridding tool provided by the Climate Data Operators (Schulzweida et al. 2019) package to re-grid the ERA5 dataset to CESM2's resolution through a second-order conservative remapping method. As with the CESM2 data, we take daily values of ERA5 precipitation and 2-meter temperature across the same spatial domains and remove the daily third-order polynomial trend in both fields to get precipitation and temperature anomalies for the entire 62-year dataset. Thereafter, we follow the same procedure as for CESM2 to label the ERA5 temperature anomalies into each of the three temperature classes based on ERA5 location-specific percentile temperature thresholds at each of the 18 locations.

Chapter 3

Interpretable Neural Network Framework

We use the interpretable architecture of ProtoLNet from Barnes et al. (2022) with the network architecture shown in Figure 3.1a. Here, we discuss the key aspects of ProtoLNet, but encourage readers to refer to Barnes et al. (2022) for a more comprehensive overview of ProtoLNet and the original ProtoPNet of Chen et al. (2018).

3.1 ProtoLNet architecture

The model begins with a base convolutional neural network (CNN), which is denoted by the green rectangle in Figure 3.1a. The base CNN has four convolutional layers of 16 kernels each of size 3x3. An average pooling layer with stride length of two and a kernel size of 2x2 follows each convolutional layer. Having four convolutional layers allows the resulting latent patches to be large enough to capture large-scale precipitation features associated with climate-scale anomalies like the MJO and ENSO. We found that using three convolutional layers resulted in prototypes that were too small to have interpretable precipitation features (not shown). We elect to pre-train the base CNN to act as a baseline for ProtoLNet. Following Barnes et al. (2022), the output from the final average pooling layer is flattened, and then fed into a final dense layer of 32 nodes that is then fed into the final output layer of 3 nodes to perform classification. The output from the final layer is converted to predictions that sum to 1.0 through a softmax activation function. Each base CNN is trained via stochastic gradient descent using the Adam optimizer with early stopping and standard cross-entropy loss. The learning rate is set to 0.00008.

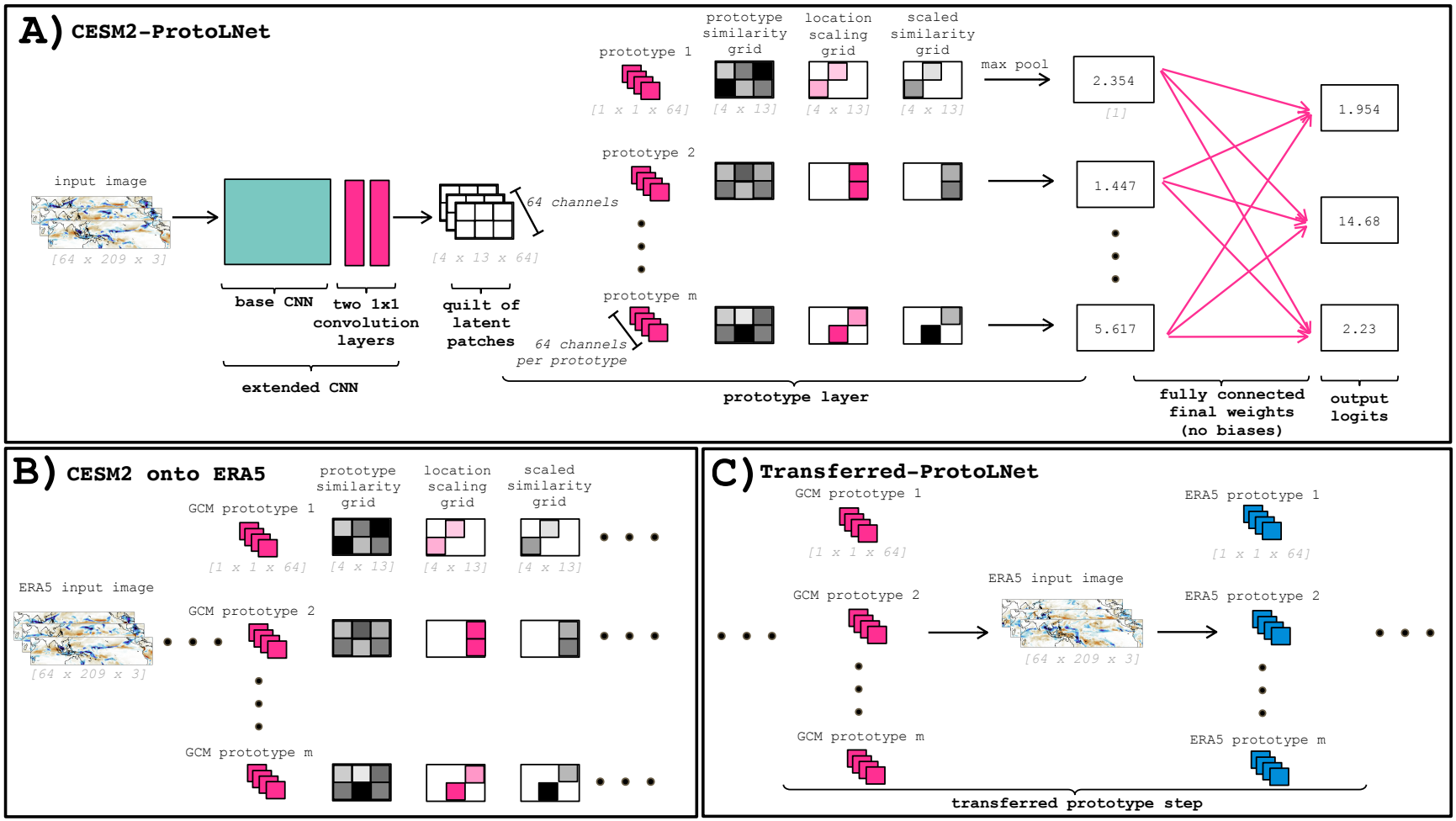

Figure 3.1: Schematic outlining the ProtoLNet architecture with additional modifications used only for ERA5 input. (a) Generalized CESM2 ProtoLNet use case with consistent tensor dimensions where the magenta components are trained by the network and the monochrome components are directly computed. (b) ERA5 images are input directly into the network using prototypes learned from CESM2 data. (c) Prototypes learned from CESM2 data are transferred to ERA5 prototypes by pushing ERA5 input directly through the learned prototype layer. This process takes a CESM2 prototype and calculates a SimilarityScore for each ERA5 input image. The ERA5 input image with the highest SimilarityScore is the new prototype.

We focus our initial results in Chapter 4 on predicting temperature anomalies over Anchorage, Alaska (Location #5 in Figure A.1), since this is an area with consistently documented MJO and ENSO teleconnections (Wang et al. 2020). We found that across all locations, including Anchorage, the base CNN is sensitive to the choice of network seed initialization. Out of 50 random initialization seeds, only 8 seeds have accuracies higher than random chance for Anchorage. Even so, we train ProtoLNet using all 50 pre-trained base CNNs since those without skill may still have learned relationships beyond those randomly initialized that are useful for training ProtoLNet.

We insert the pre-trained base CNN into the ProtoLNet architecture and train the entire network shown in Figure 3.1a. Each sample image is pushed through the extended CNN, which includes the two 1x1 convolutional layers that change the base CNN output dimensions to be consistent with the following prototype layer. This extended CNN outputs a "quilt" of latent patches belonging to the input image. The latent patches are then pushed to the prototype layer, which will ultimately provide interpretability. For each class, ProtoLNet learns a set of representative latent patches which are the prototypes for that class. In our work, ProtoLNet ingests precipitation anomalies and learns 10 prototypes for each of the three temperature classes for a total of 30 prototypes. Each prototype then receives a score (as quantified in the next section) based on how similar it looks to the sample's latent patch. The novel feature of ProtoLNet is that a location scaling grid is learned with every prototype (for a total 30 local scaling grids). The location scaling grid ensures a prototype only provides points to a class if the sample latent patch is co-located with it. Thus, this is a form of soft attention (Vaswani et al. 2017). The scores for each prototype are then connected to the output layer using a fully connected layer with trained weights, no bias, and a linear activation function. After adding up the total points provided by each prototype to each class, the class whose prototypes collectively score the highest on the sample latent patch is the predicted class.

3.2 Training ProtoLNet

Prior to training ProtoLNet, we initialized the prototypes with random values drawn from a uniform distribution (0.0-1.0). The corresponding location scaling grids are initialized with ones at every point. The final weights are initialized to -0.5 except for weights connecting a prototype to its given class, which are initialized to 1.0. ProtoLNet is then trained in three stages. Stage 1 trains the base CNN, the two 1x1 convolutional layers, the prototypes, and location scaling grid. Stage 2 replaces each stage 1 learned prototype with the training latent patch with the highest SimilarityScore from the same class. The SimilarityScore is computed as a function of the L2 norm (i.e distance) between the learned prototype and training latent patch. If the distance is low, then the SimilarityScore will be high. This process constrains ProtoLNet as it forces the network to find latent patches in the training set that may not be exact replicas of what it learned via optimization, but that look as similar as possible. However, this step makes ProtoLNet interpretable as we can exactly see which specific training samples are used as prototypes by the network. Stage 3 trains the final weights alone to lower the cross-entropy loss in the final output. We train ProtoLNet for a total of 15 stages (5 complete cycles of stages 1-3). Stages 1 and 3 are run with early stopping and a learning rate of 0.01. The learning rate is reduced by an order of magnitude at the end of each complete cycle. Further details about ProtoLNet's training process and the SimilarityScore calculation can be found in Section 2 of Barnes et al. (2022).

3.3 ERA5 applications of ProtoLNet

While CESM2-ProtoLNet is trained only on CESM2 data, we also perform inference on ERA5 observations using learned CESM2 prototypes (Figure 3.1b). We use the entire ERA5 dataset as

it is only used for inference, not training new prototypes. Thus, the trained CESM2-ProtoLNet is frozen during this process.

ProtoLNet can also be used for transfer learning on ERA5 data. Transfer learning is the process of leveraging knowledge learned by one model to train on another problem (e.g. Aytar and Zisserman 2011; Oquab et al. 2014). In our case, we use prototypes from CESM2-ProtoLNet (trained on CESM2) and transfer them to ERA5 prototypes (Figure 3.1c). ProtoLNet is designed to be able to repeat stage 2 of the training process (selection of the prototypes) without requiring the other stages. That is, all the weights in the model are frozen and only the learned prototypes are replaced with no additional learning. This replacement process functions exactly the same as in ProtoLNet, where the current prototypes are replaced by the training latent patch with the greatest SimilarityScore. Thus, CESM2-ProtoLNet ingests ERA5 precipitation anomaly fields and replaces the already learned CESM2 prototypes with their most similar ERA5 latent patch to essentially transfer the prototypes from CESM2 fields to ERA5 fields (referred to as Transferred-ProtoLNet). Thereafter, Transferred-ProtoLNet can ingest daily ERA5 precipitation anomalies and make predictions using ERA5 prototypes. By using an interpretable network like ProtoLNet, we have a method to perform transfer learning from climate model fields to observations in an interpretable way.

Chapter 4

Results

4.1 CESM2-ProtoLNet over Alaska

The accuracies of CESM2-ProtoLNet over Anchorage, Alaska (Location #5), for 50 different networks (which differ by their initialization seed and data split) are shown in Figure 4.1. These are also the same 50 initialization seeds used to train the base CNN. 27 of the 50 networks have accuracies greater than random chance, which is an increase from only 8 seeds when training the base CNN. CESM2-ProtoLNet is constrained to use prototypes directly from CESM2 input, and we hypothesize that this acts as a regularizer, allowing it to more readily learn skill for more seeds compared to the "black box" base CNN (Rudin et al. 2022). For the 27 CESM2-ProtoLNet networks that learned, accuracy increases as model confidence increases. In Figure 4.1, model confidence is defined as the softmax probability a sample belongs to a certain class, with higher values indicating higher confidence in the prediction. This type of analysis is referred to as a "discard test", as our metric "discards" samples the network is less confident in. The fact that ProtoLNet is more accurate when it is more confident suggests that it is identifying forecasts of opportunity in the region (Mayer and Barnes 2021). The dark magenta line is the accuracy of ProtoLNet for the particular network using the same seed as the dark green line from the corresponding base CNN in Figure A.2. We focus on this network throughout the rest of this paper but note that our results are robust if we had chosen other networks to highlight.

One advantage of using an interpretable architecture is the ability to dissect how the model made its prediction. Figure 4.2 shows an example cold class temperature prediction. In the context

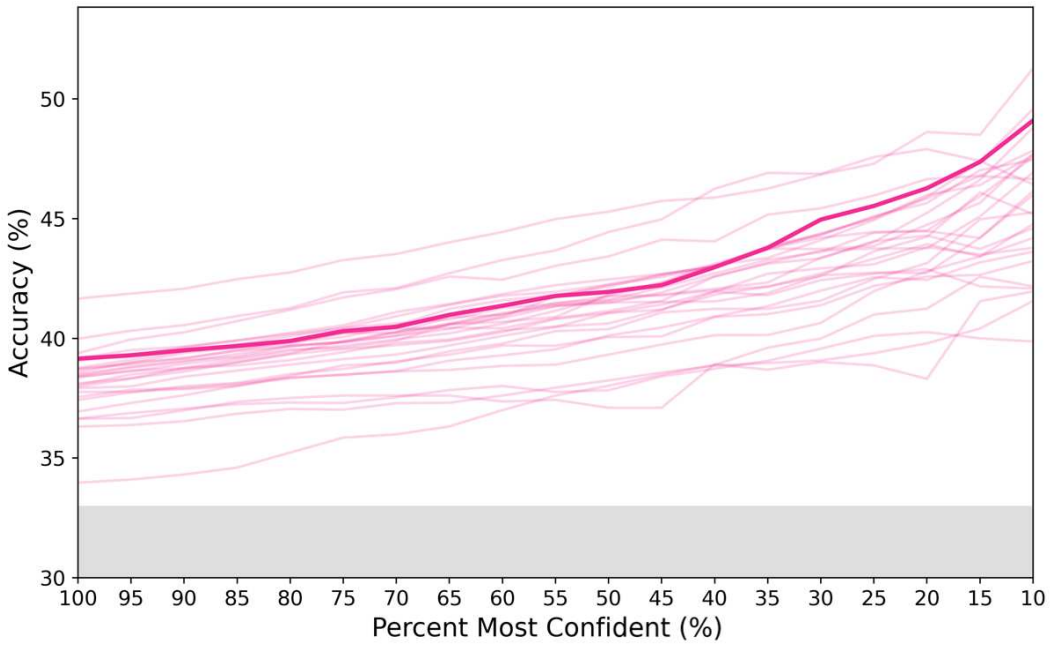

Figure 4.1: Discard test showing CESM2-ProtoLNet accuracy versus confidence for temperature class predictions near Anchorage, Alaska (Location #5). Testing data is split into three balanced classes, creating a baseline of 33.33% accuracy that is represented by the gray box. The dark magenta line is the accuracy of the selected network discussed throughout the main text. The light magenta lines are accuracies of other seeds at the same location.

of CESM2-ProtoLNet, we can interpret how the prediction is made by directly comparing the sample and the winning prototype #9 (i.e. the prototype that contributed the most points towards the learned class). For the sample (Figure 4.2a), there is a line of anomalously dry conditions stretching across the tropical Pacific Ocean with an anomalously wet area north of it, which collectively looks like a signature of a La Niña event (Philander 1985). The sample is labeled as belonging to the cold class in the upper left. The corresponding (also referred to as "winning") prototype (Figure 4.2b), also belonging to the cold class, looks very similar to the sample with the same stretch of anonymously dry conditions with an area of anonymously wet conditions north of it. This example illustrates the decision-making process of CESM2-ProtoLNet. The corresponding location scaling grid (Figure 4.2c) confirms that the prototype similarities are only weighted highly for the sample because they are located at the same longitude and longitude. That is, this prototypical feature only

adds points to the cold class if it is found over the western Pacific Ocean. The point contributions in Figure 4.2d show how the sample was correctly classified into the cold class due to the large contribution (about 1.2 points) from prototype #9 to the cold class. The prototypes from the neutral and warm class have low point contributions to their respective totals.

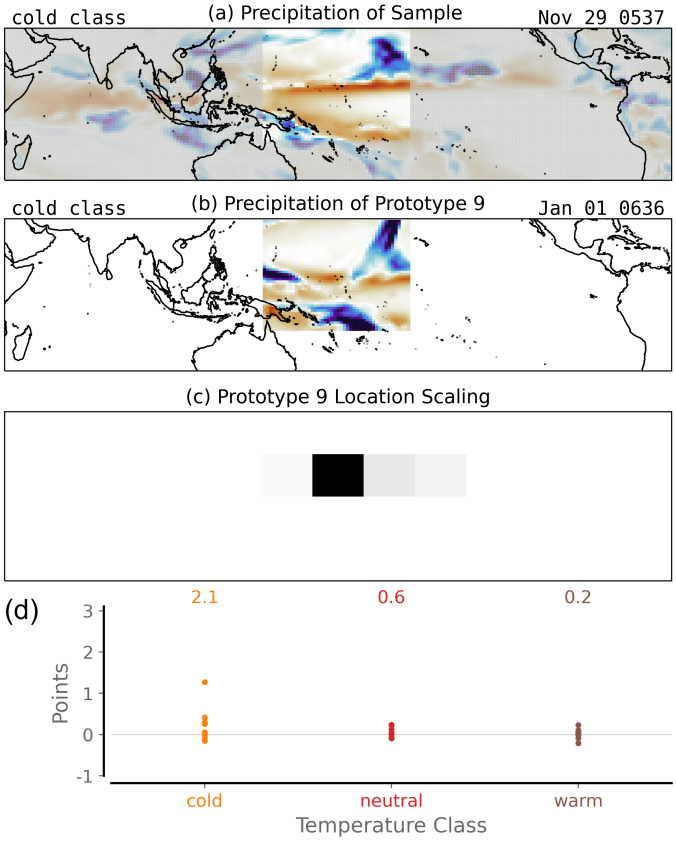

Figure 4.2: An example cold class temperature prediction from CESM2-ProtoLNet for Anchorage, Alaska (Location #5), trained on CESM2 precipitation anomalies. (a) The precipitation anomaly sample with the area outside the prototype patch muted. (b) The prototype which contributed the most amount of points towards the predicted class for the sample shown in (a). (c) The corresponding prototype location scaling grid. (d) The point contributions from each prototype, grouped together by color, to their corresponding class.

Like Figure 4.2, Figure 4.3 depicts an example prediction by CESM2-ProtoLNet, but this time for the warm class. The sample latent patch shows a long stretch of anomalously wet conditions over the Pacific Ocean with an anomalously dry area below it (Figure 4.3a). The prototype latent patch in Figure 4.3b looks extremely similar with the same general structure. The location scaling grid, shown in Figure 4.3c, reaffirms that the prototype and sample are within the same region. The point distribution in Figure 4.3d shows the warm class as the clear winner with multiple prototypes contributing 0.5 points or more. It is worth noting that, like in the cold class example, the losing classes had near-zero or negative points. Furthermore, the prototype for the warm class is almost the opposite to that of the cold class example in Figure 4.2b, which suggests that CESM2-ProtoLNet learned distinct patterns for each class.

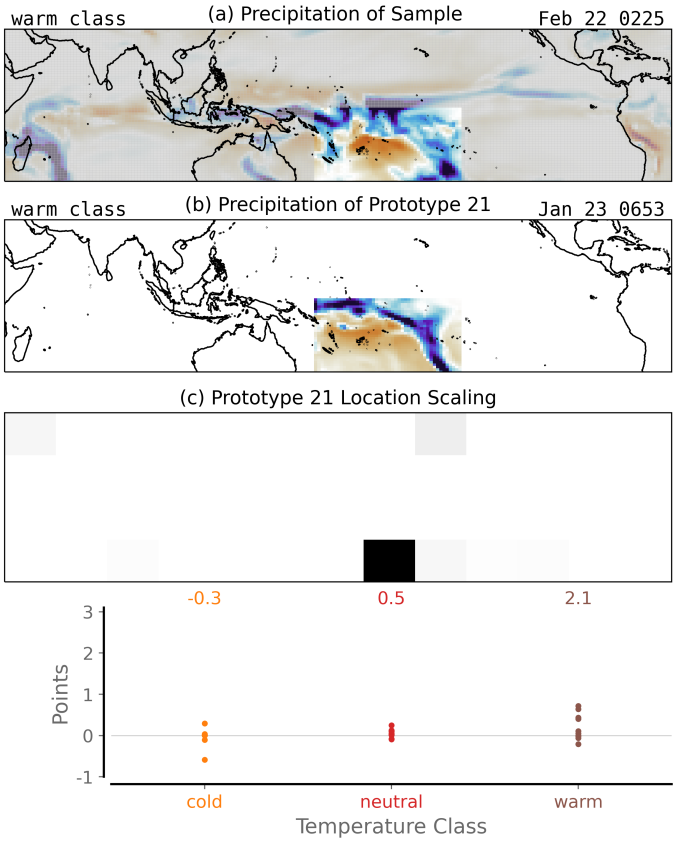

Figure 4.3: As in Figure 4.2, but for a warm class temperature prediction (Location #5).

Five of the ten cold class prototypes from CESM2-ProtoLNet are shown in Figure 4.4a. The five prototypes are listed by average correctly classified point contribution in descending order, and the date of each prototype is in the upper right. Since we use a long CESM2 control simulation, we define the years progressively starting at the year 0201. For each prototype, the network has learned certain precipitation features from specific training samples that are representative of the class. We can investigate the types of patterns CESM2-ProtoLNet has learned by looking at prototypes #4 and #8 from Figure 4.4a. Prototype #4 shows a stretch of anomalously dry conditions across the central tropical Pacific Ocean with another stretch of anomalously wet conditions situated poleward. Like prototype #9 in Figure 4.2b, this is a La Niña-like feature (Philander 1985),

which means samples being classified into the cold class with a high point contribution from prototypes #4 or #9 must exhibit La-Niña like features. For prototype #8, even though it appears that there are no strong anomalies in the prototype, CESM2-ProtoLNet can learn relationships related to a lack of strong anomalous features. The location scaling grid in Figure 4.4b shows that both prototypes are most important over the Pacific. The Pacific Ocean is an expected region to find ENSO precipitation features, and the interpretable design of CESM2-ProtoLNet allows us to visualize exactly which anomalies and their locations are important for accurate S2S predictions.

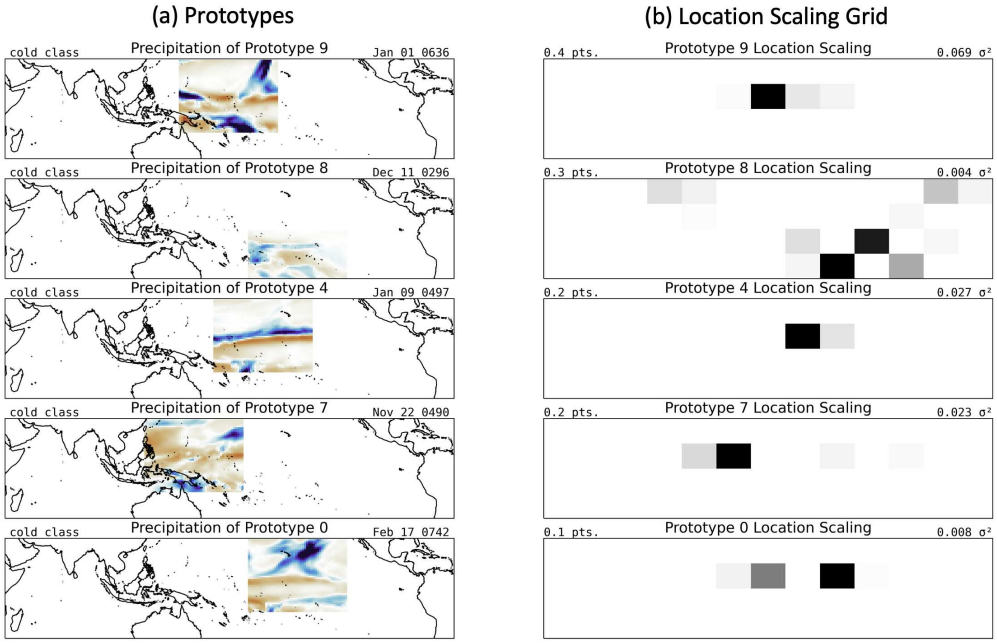

Figure 4.4: (a) Five CESM2 precipitation prototypes learned by CESM2-ProtoLNet for the cold temperature class prediction near Anchorage, Alaska (Location #5). The blue values represent positive precipitation anomalies while orange values denote negative anomalies. (b) The associated location scaling grid for each prototype, where darker values correspond to more importance placed on the specific location. The prototypes are ordered by average point contribution towards the class prediction (given in the upper left of each panel). The variance of the points contributed to each prediction by each prototype is represented by σ^2, which is shown in the upper right of each panel.

We have found that when a prediction is made, often only three or four prototypes contribute the majority of points towards the final predicted class score. Thus, it is easy for prototypes, which on average contribute less, to be dismissed even though they may be identifying important patterns. This may be particularly important for our S2S application as prototypes that contribute greater amounts of points on a subset of samples may be finding forecasts of opportunity. We identify prototypes that contribute greater amounts of points on a subset of samples via each prototype's "variance score". The variance score, shown in the upper right of Figure 4.4b as σ^2, is a measure of the variance across all of the points a prototype contributes to a correct classification. Prototypes with a low variance score, like prototype #0 in Figure 4.4, contribute nearly the same amount of points to every correct classification. In contrast, prototype #9 has a much larger variance score, indicating that prototype #9 may be useful for distinguishing same-class samples from one another and thus for distinguishing forecasts of opportunity.

Figure 4.5a shows the top 3 cold class prototypes based on variance score. Figure 4.5b shows the composite of precipitation anomalies for samples when the prototype contributed the most amount of points to a correct classification. In all three composites, we can see La Niña-like signatures with various subtle differences between them. Previous work has shown that La Niña is linked to cold temperature anomalies over Alaska (Papineau 2001). Thus, ProtoLNet has learned that ENSO and MJO signals are also important in finding forecasts of opportunities.

4.2 Predictions along the western coast of North America

We train CESM2-ProtoLNet (using the same seed and data split as earlier) to make predictions for 17 other locations along the western coast of North America (Figure 4.6). These locations are all within areas with documented MJO and ENSO S2S teleconnections (Arcodia et al. 2023).

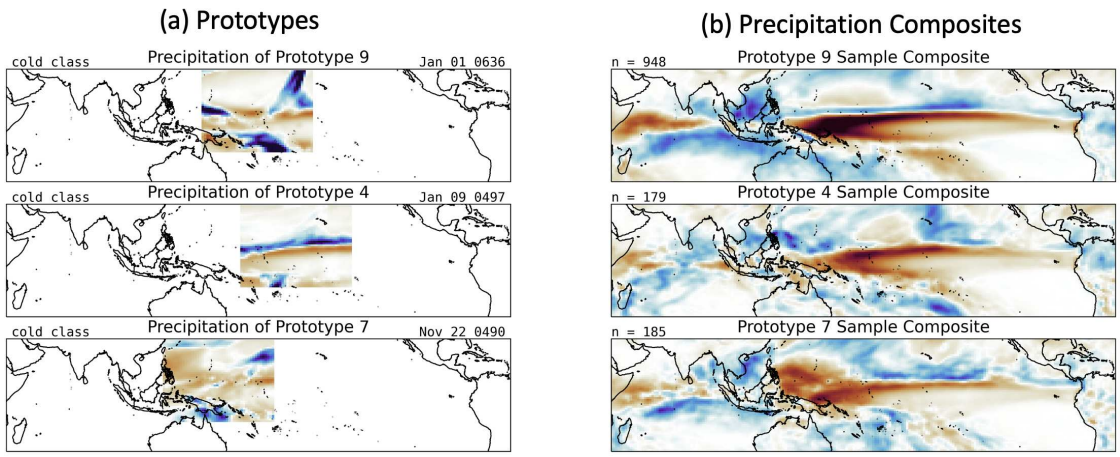

Figure 4.5: (a) Cold class CESM2 prototypes trained on CESM2 precipitation anomalies. (b) CESM2 precipitation composite of samples whose corresponding prototype was the highest scoring. The number of samples used is given in the upper left.

To identify forecasts of opportunity along the coast, we compute the accuracy of each network for both the 100% and 20% of the most confident samples for each location. We find that 12 out of 17 locations (using the same random seed) exhibit some skill above random chance when using the entire testing set. For each of these 12 locations, the skill increases when only using the 20% most confident predictions. Thus, CESM2-ProtoLNet has found forecasts of opportunity along the western coast of North America. Forecasts of opportunity using tropical Pacific Ocean precipitation on the western coast of North America have also been documented in other work using black-box models (e.g. Arcodia et al. 2023). The prototypical precipitation patterns found by the prototypes (not shown) are also consistent along the coast. That is, cold class prototypes mainly show La Niña-like patterns and the warm class show El Niño-like patterns. While there may be other phenomena, like the MJO, contributing towards the prediction, it is likely that the ENSO signals are more pronounced and thus showing up more in the prototypes.

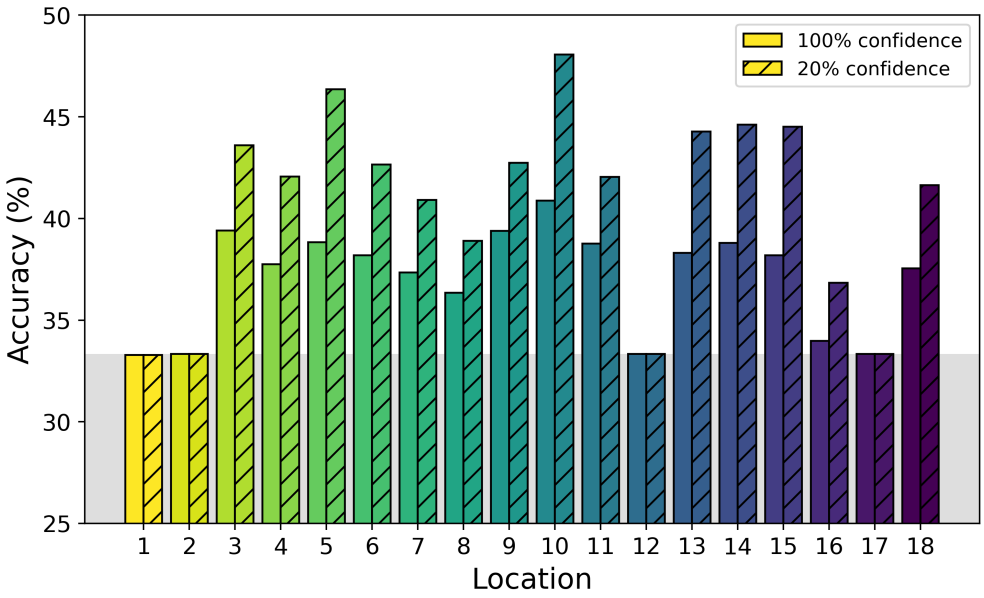

Figure 4.6: CESM2 testing accuracy of CESM2-ProtoLNet trained on CESM2 precipitation anomalies across locations on the western coast of North America (see Figure A.1). The solid color bars represent accuracies calculated using the entire testing set while the hatched bars show the accuracy using only the 20% most confident testing predictions for that location. Grey shading denotes accuracies at or below random chance.

4.3 Predictions using ERA5

Thus far, we have focused on evaluating CESM2-ProtoLNet using a CESM2 testing set. Although CESM2-ProtoLNet was trained on CESM2 data, we can use it to perform inference on ERA5 observations in two ways. The first way is using CESM2-ProtoLNet prototypes directly on ERA5 input to make predictions (CESM2 onto ERA5). The second way is to re-project the prototypes learned by CESM2-ProtoLNet onto ERA5 data and use the resulting ERA5 prototypes to make predictions on ERA5 input (Transferred-ProtoLNet). We directly compare the accuracies of our three experimental set-ups via their respective discard test accuracies in Figure 4.7. CESM2-ProtoLNet (pink) has the highest overall accuracy at all confidences, which is reasonable considering the testing data and training data used to compute the prototypes are from the same data distribution. Using CESM2-ProtoLNet prototypes on ERA5 input (orange) exhibits lower accuracies (but well above random chance), and the accuracies still increase as confidence increases, further demonstrating how prototypes learned by CESM2-ProtoLNet are useful for finding forecasts of opportunity within the observations. Since ERA5 input may not have the exact same physics and dynamics found on CESM2 latent patches, high accuracy further shows that CESM2-ProtoLNet found general patterns applicable to both datasets.

For this portion of the study, we show results from all three set-ups using a different initialization seed than the results in Sections 4.1 and 4.2. Transferred-ProtoLNet (blue), exhibits accuracies very close to the accuracy of using CESM2-ProtoLNet prototypes on ERA5 input (orange). When transferring the prototypes from CESM2 to ERA5, there is a chance that the resulting transferred prototypes are different from the original CESM2 prototypes. For example, precipitation features and gradients could be lost or different features could be introduced. One way this could happen is

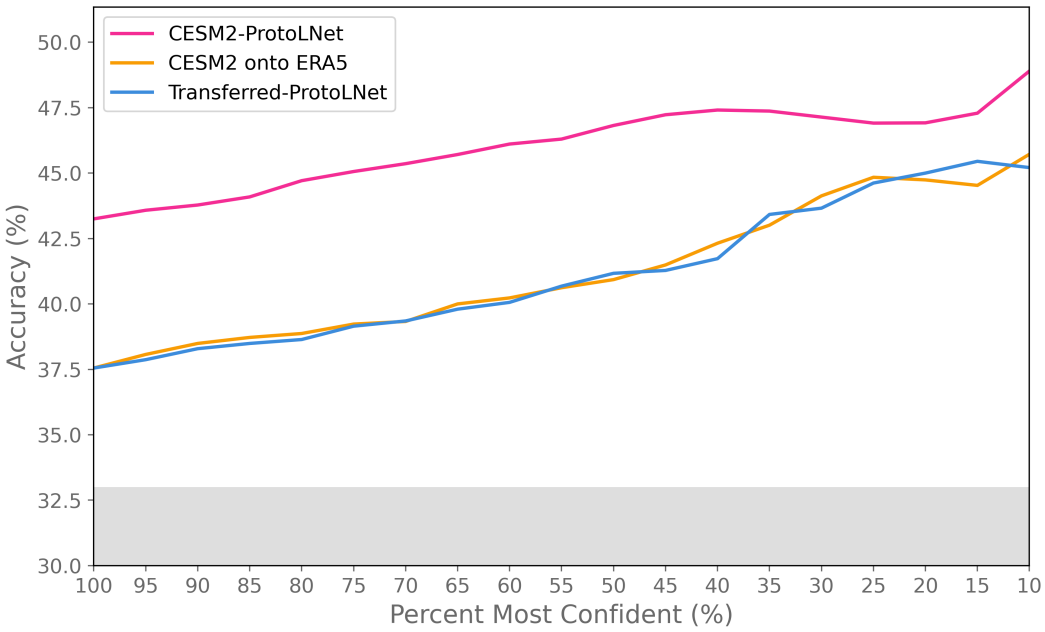

Figure 4.7: Discard test showing accuracies from three different ProtoLNet prediction methods using ProtoLNet over Vancouver, Canada (Location #12 in Figure A.1). The magenta line represents predictions from ProtoLNet on CESM data (CESM2-ProtoLNet). The orange line represents predictions of ERA5 using CESM2 prototypes (CESM2 onto ERA5). The blue line represents predictions of ERA5 using transferred ERA5 prototypes (Transferred-ProtoLNet).

through the model biases of CESM2. CESM2 shifts ENSO spatial patterns, such as SST anomalies, more westward compared to observations (Capotondi et al. 2020). Thus, this would affect precipitation anomalies in the region which could result in patterns in CESM2 prototypes that are not found in ERA5 data (which has its own biases (Lavers et al. 2022)) . Therefore, the resulting ERA5 prototypes could lack important features leading to lower skill. Another potential issue is the volume of available ERA5 data. Since there are only 61 years of ERA5 data compared to the 350 years of CESM training data, finding an exact replica of each prototype with only a sixth of the data may be difficult, even in the absence of climate biases within CESM2. Despite these potential issues, the high accuracy from Transferred-ProtoLNet shows that the relationships identified by ProtoLNet are robust.

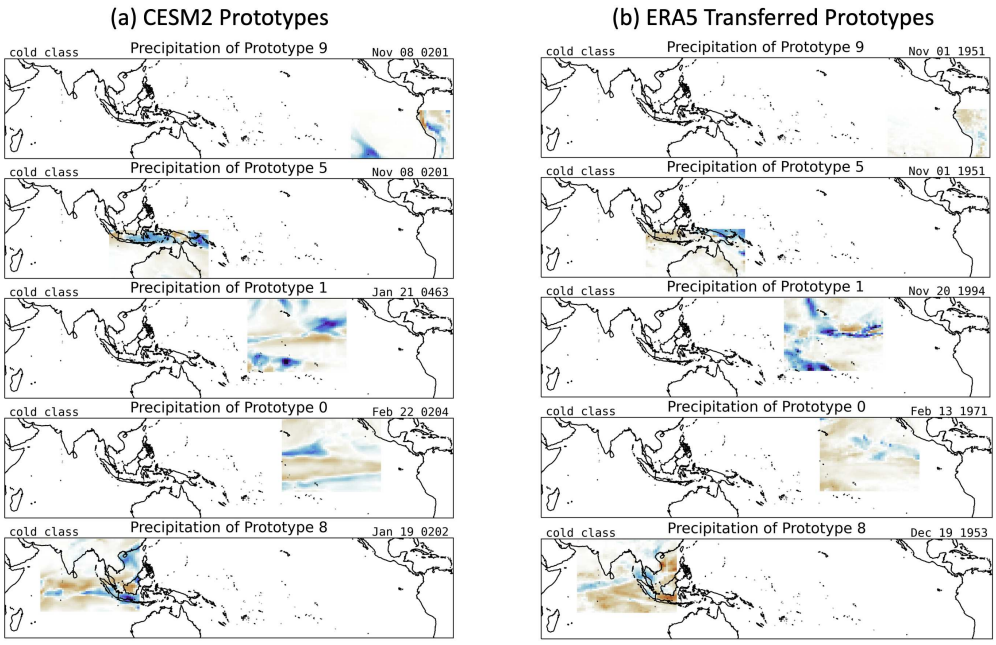

Figure 4.8: (a) Cold class CESM2 prototypes trained on CESM2 precipitation anomalies. (b) Corresponding ERA5 prototypes obtained by transferring the CESM2 prototypes to the ERA5 precipitation field. Historical dates for the ERA5 prototypes are given in the upper right hand corner of each panel. Blue shading represents positive precipitation anomalies while orange shading denotes negative anomalies. Both sets of prototypes are from Location #13 (near Vancouver, Canada).

We can visualize how CESM2 prototypes (Figure 4.8a) are transferred to ERA5 prototypes (Figure 4.8b) by directly comparing them. Once again, this is straightforward due to the interpretable architecture of ProtoLNet. We display results for Location #13 from Figure A.3 as it is the location with the highest overall accuracy for CESM2-ProtoLNet but conclusions are robust for other locations as well. CESM2 prototype #1, exhibits anomalously wet conditions near the bottom left and upper right corners of the region with a dry area in between. Similarly, Transferred-ProtoLNet finds the most similar ERA5 latent patch to CESM2 prototype #1, which is shown in Figure 4.8b. The Transferred-ProtoLNet prototype #1 from ERA5 has a very similar structure with minor differences related to more anomalously wet conditions near the middle of the region. As discussed earlier, we do not expect the prototypes to look identical. Nevertheless, despite these data limitations, the prototypes share many similar features.

One particularly novel aspect of Transferred-ProtoLNet is that we obtain real-world dates attached to each prototype. For example in Figure 4.8a, CESM2-ProtoLNet prototype #0 has a date of 22 February 0204. The corresponding Transferred-ProtoLNet prototype #0 in Figure 4.8b has a real-world, observed date of 13 February 1971. This is notable as the conditions in the tropics from 13 February 1971 have been identified as prototypically conducive to teleconnections that lead to cold temperature anomalies over Vancouver, Canada. That is, this one precipitation event is helpful in predicting temperature for many different years and events both in the past and future (relative to 13 February 1971). It is also possible to examine the conditions for this date further to look for possible signals and events outside of just the prototype region that are helpful in identifying teleconnections. Similarly, examining the dates before and after 13 February 1971 could be helpful in understanding the prototype. We leave this for future work.

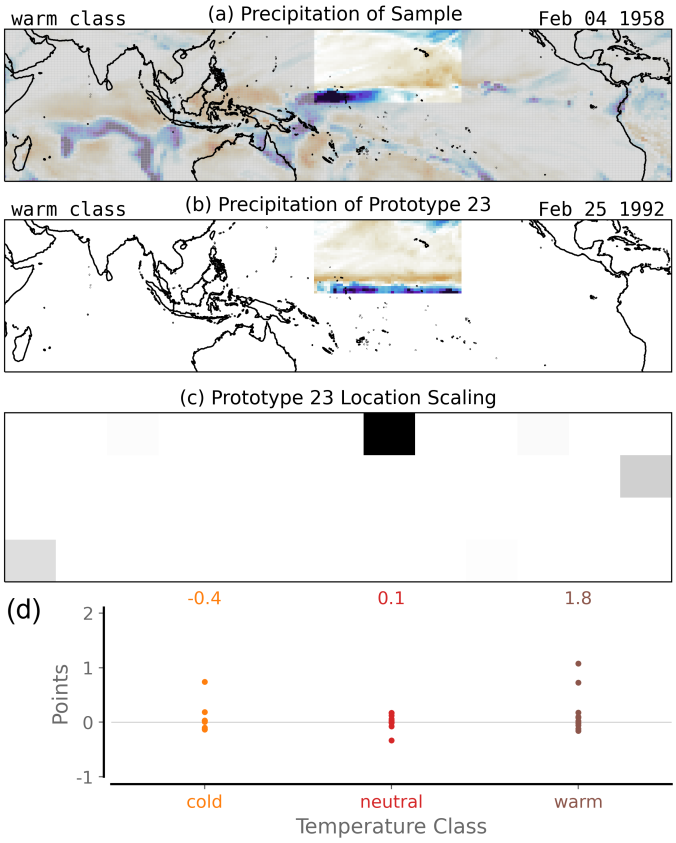

Figure 4.9: As in Figure 4.2, but for predictions of the ERA5 warm class using Transferred-ProtoLNet with ERA5 precipitation anomalies and transferred prototypes.

As we did for CESM2-ProtoLNet, we can take advantage of Transferred-ProtoLNet's interpretable architecture to dissect how the network's predictions are made. Figure 4.9 shows an example warm class prediction by Transferred-ProtoLNet (using real-world prototypes). The sample precipitation anomaly field, which occurred on 16 February 1958 (Figure 4.9a), shows a small stretch of anomalously wet conditions near the bottom of the latent patch with anomalously dry conditions poleward. The corresponding prototype in Figure 4.9b shows a very similar setup that occurred on 25 February 1992. Transferred-ProtoLNet has successfully taken a CESM2 prototype, found a real-world date to represent it, and then made a prediction for another real-world date using a prototype which originally stemmed from climate model data. There are apparent similarities with this prediction and a warm class prediction made by CESM2-ProtoLNet in Figure 4.3. Both show a stretch of anomalously wet conditions in their prototypes along with high warm class scores and negative to near-zero scores for the cold and neutral class (Figure 4.9d).

Using Transferred-ProtoLNet, we compare the accuracies of all 18 locations along the western coast of North America in ERA5 (Figure 4.10). The accuracies are overall slightly lower than CESM2-ProtoLNet's (trained using the same seed; compare to Figure A.3), which is not unexpected as the network was originally trained on CESM2 data. Instead, it is more surprising that this works at all! However, recent work by Mayer and Barnes (2022) have also demonstrated that neural networks trained on CESM2 data can exhibit skill when applied to ERA5 S2S prediction tasks. Many of the locations still have increasing accuracy with confidence, which demonstrates that forecasts of opportunity are present and identified in ERA5. Moreover, all the locations which had skill in CESM2-ProtoLNet (Figure A.3) continue to have skill in Transferred-ProtoLNet. Thus, the spatial robustness of the patterns learned by CESM2-ProtoLNet continued to be applicable to ERA5 data.

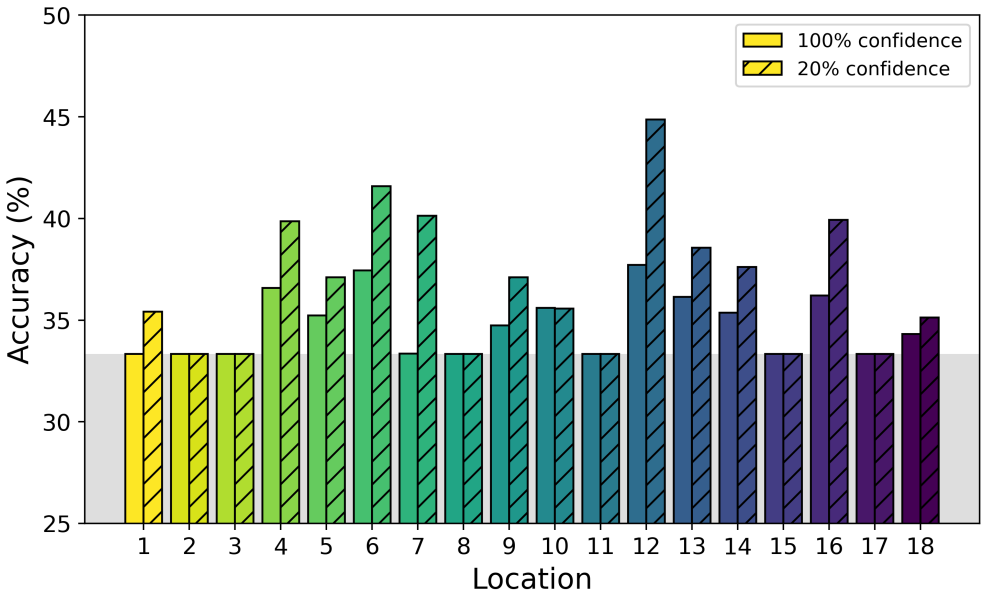

Figure 4.10: ERA5 testing accuracy of ProtoLNet using transferred ERA5 prototypes to predict ERA5 temperature anomalies across locations on the western coast of North America (Figure A.1). The solid color bars represent accuracies calculated using the entire testing set while the hatched bars show the accuracy using only the 20% most confident predictions for that location. Grey shading denotes accuracies at below random chance.

Chapter 5

Discussion and Conclusions

As the proliferation of neural networks increases in geoscience research, finding ways to understand a network's decision-making process and gauge trust has become paramount (Bostrom et al. 2023). Using networks like ProtoLNet, whose architectures are designed to be inherently interpretable, is one way to understand the model's exact decision-making process in a way that is comprehensible to humans (Rudin et al. 2022). The interpretability of ProtoLNet arises from the use of prototypes, which allow for a comprehensible representation of latent patches. Here, we apply ProtoLNet to a subseasonal-to-seasonal (S2S) prediction task. Since tropical variability is well-known to be a dominant driver of midlatitude S2S prediction skill, we use tropical precipitation as an input into ProtoLNet with an expectation that the network will identify regional-scale patterns associated with the MJO and ENSO as prototypes. In doing this, CESM2-ProtoLNet, using CESM2 as the input, was able to pick up on prototypical precipitation anomaly patterns, which it used to make skillful temperature predictions. These patterns are representative of MJO and ENSO features, which suggests that CESM2-ProtoLNet is able to identify these features. Thus, we are able to understand how ProtoLNet made its prediction by looking at the samples and prototypes ourselves and finding the similarities between them, which is exactly the same process that ProtoLNet employs.

Interpretable networks have often been falsely characterized as requiring a trade-off in skill for interpretability. However, we find that CESM2-ProtoLNet exhibits skill well above random chance and on par with a black box CNN. Furthermore, the accuracy of CESM2-ProtoLNet increases

as confidence increases, demonstrating that CESM-ProtoLNet identifies forecasts of opportunity across much of the western coast of North America.

We show that ProtoLNet can be extended to predict on ERA5 observations following two different approaches. First, using CESM2-ProtoLNet on ERA5 observations directly, we show that the skill is above random chance and accuracy increases as confidence does. This demonstrates that the prototypical patterns learned from CESM2 can be used for skillful ERA5 predictions. Likewise, the prototypes have patterns that specifically distinguish forecasts of opportunities. Second, we applied transfer learning by projecting CESM2-Prototypes onto ERA5 fields to get ERA5 prototypes (Transferred-ProtoLNet). In doing this, each prototype now belongs to a real-world event. Future research surrounding these events can help in understanding how the conditions for each prototype lends itself in being a key prototypical pattern and if they can be used to identify more prototypes. We also find that the forecasts of opportunity found in CESM2-ProtoLNet are also found in Transferred-ProtoLNet.

While we use ProtoLNet to classify temperature across North America using only precipitation, prior research has shown that other geophysical fields, such as geopotential height, are also helpful in S2S prediction problems (e.g. Henderson et al. 2017). ProtoLNet is built to ingest multiple channels; thus, it would be straightforward to extend this study to use both precipitation and geopotential height (or any suitable combination of fields) as inputs to further explore how predictions improve and why. Each field would receive its own set of prototypes, but, ProtoLNet is currently limited in that each set of prototypes (e.g, each field's prototype #0) must have the same location scaling grid. Therefore, future work should include expanding ProtoLNet to handle multiple location scaling grids for prototype sets.

As machine learning continues to become a prominent part of the scientific process, the need to understand the decision-making of black box models will only increase. While building interpretable models is one path forward, it is important to be cognizant of the challenges that come with it, such as taking substantially more time and effort to design and implement than most black box counterparts, and requiring extensive domain knowledge to ensure interpretability from start to finish. ProtoLNet is an example of how powerful interpretability can be, and since it processes and uses latent patches from 2D or 3D inputs, it is likely applicable to any geoscience problem that can rely on prototypical-parts to make predictions.

Bibliography

Arcodia, M. C., E. A. Barnes, K. J. Mayer, J. Lee, A. Ordonez, and M.-S. Ahn, 2023: Assessing decadal variability of subseasonal forecasts of opportunity using explainable AI. *Environ. Res.: Climate*, **2 (4)**, 045 002.

Arcodia, M. C., B. P. Kirtman, and L. S. P. Siqueira, 2020: How MJO teleconnections and ENSO interference impacts U.S. precipitation. *J. Clim.*, **33 (11)**, 4621–4640.

Aytar, Y., and A. Zisserman, 2011: Tabula rasa: Model transfer for object category detection. *2011 International Conference on Computer Vision*, IEEE, 2252–2259.

Barnes, E. A., R. J. Barnes, Z. K. Martin, and J. K. Rader, 2022: This looks like that there: Interpretable neural networks for image tasks when location matters. *Artificial Intelligence for the Earth Systems*, **1 (3)**.

Bostrom, A., and Coauthors, 2023: Trust and trustworthy artificial intelligence: A research agenda for AI in the environmental sciences. *Risk Anal.*

Capotondi, A., C. Deser, A. S. Phillips, Y. Okumura, and S. M. Larson, 2020: ENSO and pacific decadal variability in the community earth system model version 2. *J. Adv. Model. Earth Syst.*, **12 (12)**.

Chen, C., O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, 2018: This looks like that: Deep learning for interpretable image recognition. 1806.10574.

Connolly, C., E. A. Barnes, P. Hassanzadeh, and M. Pritchard, 2023: Using neural networks to learn the jet stream forced response from natural variability. *Artificial Intelligence for the Earth Systems*, **2 (2)**.

Danabasoglu, G., and Coauthors, 2020: The community earth system model version 2 (CESM2). *J. Adv. Model. Earth Syst.*, **12 (2)**.

Gordon, E. M., E. A. Barnes, and J. W. Hurrell, 2021: Oceanic harbingers of pacific decadal oscillation predictability in CESM2 detected by neural networks. *Geophys. Res. Lett.*, **48 (21)**.

Ham, Y.-G., J.-H. Kim, and J.-J. Luo, 2019: Deep learning for multi-year ENSO forecasts. *Nature*, **573 (7775)**, 568–572.

Henderson, S. A., E. D. Maloney, and S.-W. Son, 2017: Madden–Julian oscillation pacific teleconnections: The impact of the basic state and MJO representation in general circulation models. *J. Clim.*, **30 (12)**, 4567–4587.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146 (730)**, 1999–2049.

Hilburn, K. A., 2023: Understanding spatial context in convolutional neural networks using explainable methods: Application to interpretable GREMLIN. *Artificial Intelligence for the Earth Systems*, **2 (3)**.

Huang, H., C. M. Patricola, E. Bercos-Hickey, Y. Zhou, A. Rhoades, M. D. Risser, and W. D. Collins, 2021: Sources of subseasonal-to-seasonal predictability of atmospheric rivers and precipitation in the western united states. *Journal of Geophysical Research: Atmospheres*, **126 (6)**, https://doi.org/10.1029/2020jd034053, URL http://dx.doi.org/10.1029/2020JD034053.

Lavers, D. A., A. Simmons, F. Vamborg, and M. J. Rodwell, 2022: An evaluation of ERA5 precipitation for climate monitoring. *Quart. J. Roy. Meteor. Soc.*, **148 (748)**, 3152–3165.

Madden, R. A., and P. R. Julian, 1971: Detection of a 40–50 day oscillation in the zonal wind in the tropical pacific. *J. Atmos. Sci.*, **28 (5)**, 702–708.

Madden, R. A., and P. R. Julian, 1972: Description of Global-Scale circulation cells in the tropics with a 40–50 day period. *J. Atmos. Sci.*, **29 (6)**, 1109–1123.

Mamalakis, A., E. A. Barnes, and I. Ebert-Uphoff, 2022: Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *Artificial Intelligence for the Earth Systems*, **1 (4)**.

Mariotti, A., and Coauthors, 2020: Forecasts of opportunity: Opening windows of skill, subseasonal and beyond. *Bull. Am. Meteorol. Soc.*, **101 (7)**, 597–601.

Mayer, K. J., and E. A. Barnes, 2021: Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophys. Res. Lett.*, **48 (10)**.

Mayer, K. J., and E. A. Barnes, 2022: Quantifying the effect of climate change on midlatitude subseasonal prediction skill provided by the tropics. *Geophys. Res. Lett.*, **49 (14)**.

McGovern, A., I. Ebert-Uphoff, D. J. Gagne, and A. Bostrom, 2022: Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. *Environmental Data Science*, **1**, e6.

Molina, M. J., and Coauthors, 2023: A review of recent and emerging machine learning applications for climate variability and weather phenomena. *Artificial Intelligence for the Earth Systems*, **2 (4)**, https://doi.org/10.1175/aies-d-22-0086.1, URL http://dx.doi.org/10.1175/AIES-D-22-0086.1.

Oquab, M., L. Bottou, I. Laptev, and J. Sivic, 2014: Learning and transferring mid-level image representations using convolutional neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 1717–1724.

Papineau, J. M., 2001: Wintertime temperature anomalies in alaska correlated with ENSO and PDO. *Int. J. Climatol.*, **21 (13)**, 1577–1592.

Philander, S. G. H., 1985: El niño and la niña. *J. Atmos. Sci.*, **42 (23)**, 2652–2662.

Rudin, C., 2019: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, **1 (5)**, 206–215.

Rudin, C., C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, 2022: Interpretable machine learning: Fundamental principles and 10 grand challenges. *ssu*, **16 (none)**, 1–85.

Schulzweida, U., L. Kornblueh, and R. Quast, 2019: CDO user guide.

Son, S.-W., H. Kim, K. Song, S.-W. Kim, P. Martineau, Y.-K. Hyun, and Y. Kim, 2020: Extratropical prediction skill of the subseasonal-to-seasonal (S2S) prediction models. *J. Geophys. Res.*, **125 (4)**.

Stan, C., D. M. Straus, J. S. Frederiksen, H. Lin, E. D. Maloney, and C. Schumacher, 2017: Review of tropical-extratropical teleconnections on intraseasonal time scales. *Rev. Geophys.*, **55 (4)**, 902–937.

Toms, B. A., E. A. Barnes, and I. Ebert-Uphoff, 2020: Physically interpretable neural networks for the geosciences: Applications to earth system variability. *J. Adv. Model. Earth Syst.*, **12 (9)**.

Trenberth, K. E., 1997: The definition of el niño. *Bull. Am. Meteorol. Soc.*, **78 (12)**, 2771–2777.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, 2017: Attention is all you need. *Adv. Neural Inf. Process. Syst.*, **30**.

Vitart, F., and Coauthors, 2017: The subseasonal to seasonal (S2S) prediction project database. *Bull. Am. Meteorol. Soc.*, **98 (1)**, 163–173.

Wang, J., H. Kim, D. Kim, S. A. Henderson, C. Stan, and E. D. Maloney, 2020: MJO teleconnections over the PNA region in climate models. part II: Impacts of the MJO and basic state. *J. Clim.*, **33 (12)**, 5081–5101.

White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorol. Appl.*, **24 (3)**, 315–325.

Appendix A

Supplementary Figures

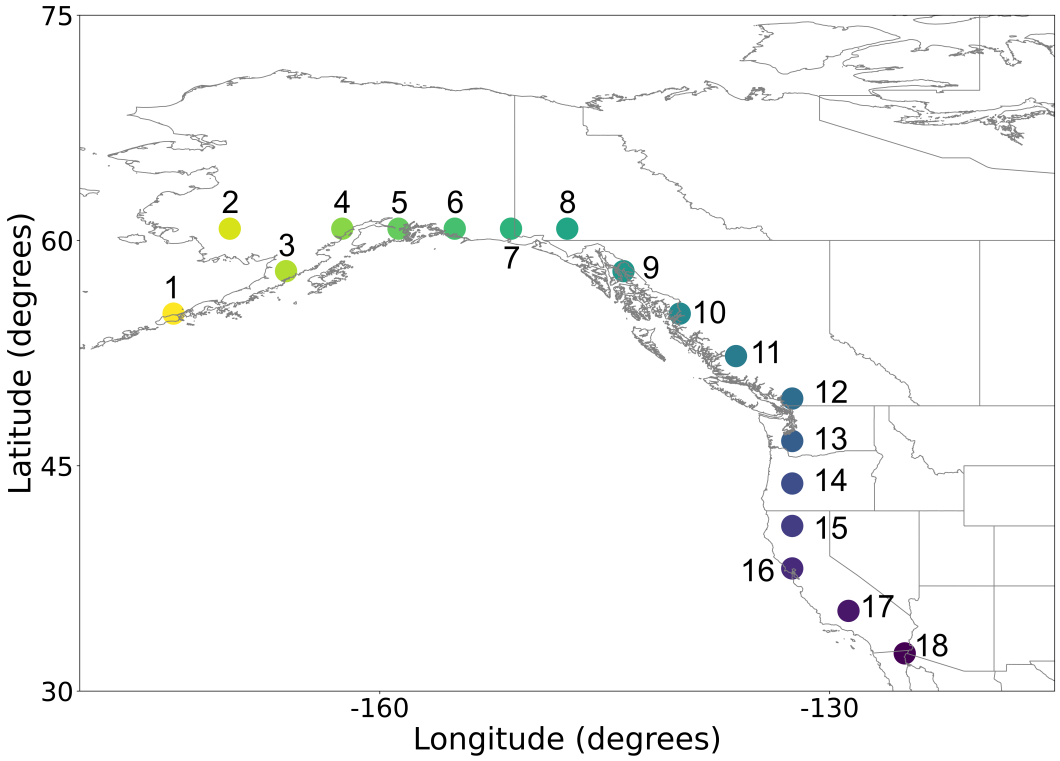

Figure A.1: Western coast of North America with colored points denoting locations for which models were trained to predict the temperature class.

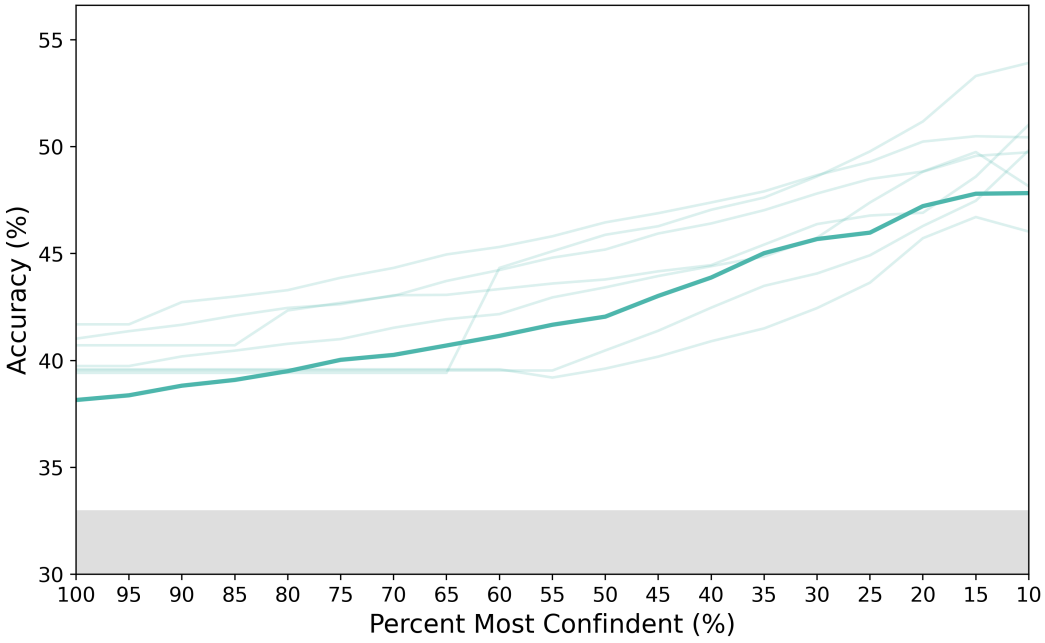

Figure A.2: Discard test showing base CNN validation accuracy versus confidence for temperature class predictions near Anchorage, Alaska (Location #5). Testing data is split into three balanced classes creating a baseline of 33.33% accuracy which is represented by the gray box. The dark green line is the accuracy of the selected network discussed throughout the main text. The light green lines are accuracies of other seeds at the same location.

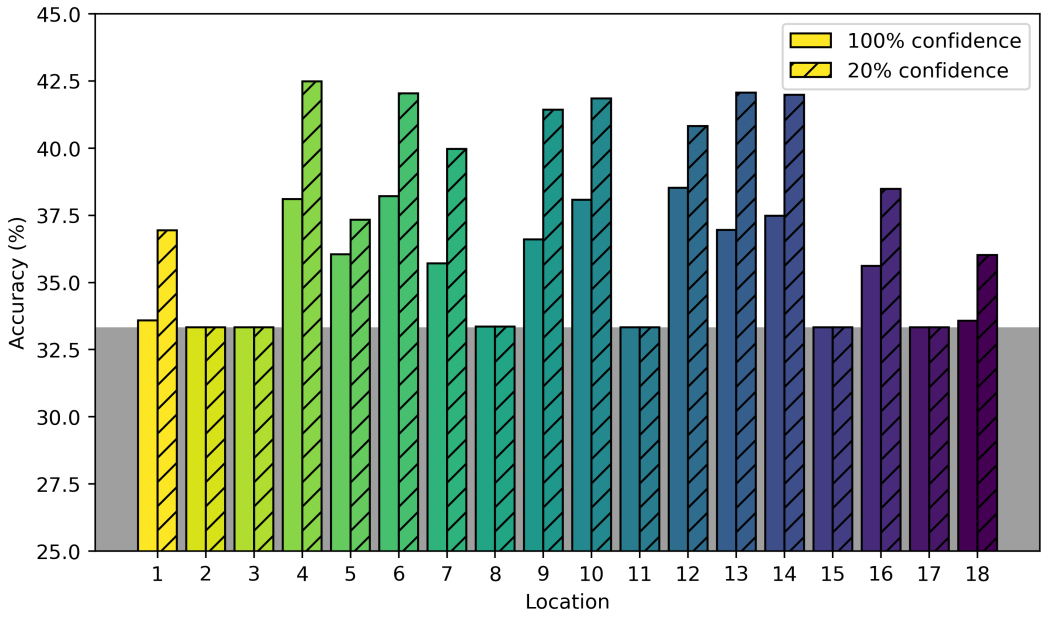

Figure A.3: As in Figure A.3, but using the same seed initialization used for Transferred-ProtoLNet.