Understanding the WPC Cluster Analysis Tools

Jonathan Rutz, Michael Staudenmaier, Matt Jeglum, and Bill Lamberson

Updated 12 May, 2022

Goal / Purpose

This document is a guide to understanding and interpreting the WPC Cluster Analysis tools available for Days 3-7 (here) and Days 8-10 (here). These tools are incredibly useful for understanding the sources of ensemble uncertainty, identifying extremes, and evaluating potential forecast scenarios via ensemble clusters. The cluster analysis groups 90 ensemble members from the EPS (50), GEFS (30) and CMCE (20) into four clusters based on similarities and dissimilarities amongst these members (more information on this clustering process is provided later). Skilled use of these tools improves our own internal understanding of potential outcomes and increases our ability to provide top-level DSS.

Section 1 is a quick guide, or summary of basic information, that can be used as a "refresher" by forecasters who have already familiarized themselves with basic training throughout the document. Section 2 provides the majority of this basic training. Section 3 addresses additional topics that come up frequently. A thorough understanding of both Sections 1 and 2 is necessary to use the WPC Cluster Analysis tools effectively.

Contents:

1. Quick Guide / Summary of Basic Information

The 00z (12z) runs are usually available online by 10z (22z).

1a. Cluster Analysis Basics

- The dominant patterns in the ensemble forecast are derived by calculating the first and second Empirical Orthogonal Functions (EOFs) of the ensemble members' 500-mb height patterns.
- The first EOF always explains the greatest percentage of spread in the 500-mb height field; the second EOF explains the second greatest percentage of spread.
- In general, an EOF will take on one of two patterns:
    1. A **dipole** centered on the ensemble-mean location of the ridge or trough. This **indicates position and/or timing uncertainty in the location of the ridge or trough** among the ensemble members.
    2. A **monopole** centered near the ensemble-mean location of the ridge or trough. This **indicates uncertainty in the amplitude of the ridge or trough** among the ensemble members.
    3. The sign (and color) of the dipole does not really matter. It just helps us identify which members look similar (positive) or dissimilar (negative) to that EOF.
- A phase space of forecast scenarios is constructed from the first two EOFs and each ensemble member is plotted on the phase space diagram using its principal components of EOF1 and EOF2 as an X and Y coordinate.
    - An ensemble member with a large principal component for a given EOF will strongly resemble the forecast scenario represented by that EOF pattern.
- The ensemble members are clustered based on where they fall in the phase space.

1b. Identifying Forecast Scenarios

1. Identify a forecast event and forecast time of interest
2. Identify the geographic region that best captures the event
3. Open tabs for "500 mb EOF Patterns", "Cluster Phase Space", and "Cluster 500 Heights" for the forecast day and region of interest.
    a. Use the EOF patterns to identify regions of uncertainty amongst ensemble members
    b. Confirm for yourself the relationships between the EOFs, ensemble members plotted on the phase space diagram, and the 500-mb height clusters
4. Open tabs for "Cluster 500 Heights", "Maximum Temperatures", "Minimum Temperatures", and "24- or 72-h QPF" for the forecast day and region of interest.
    a. Use these clusters to assess possible forecast scenarios

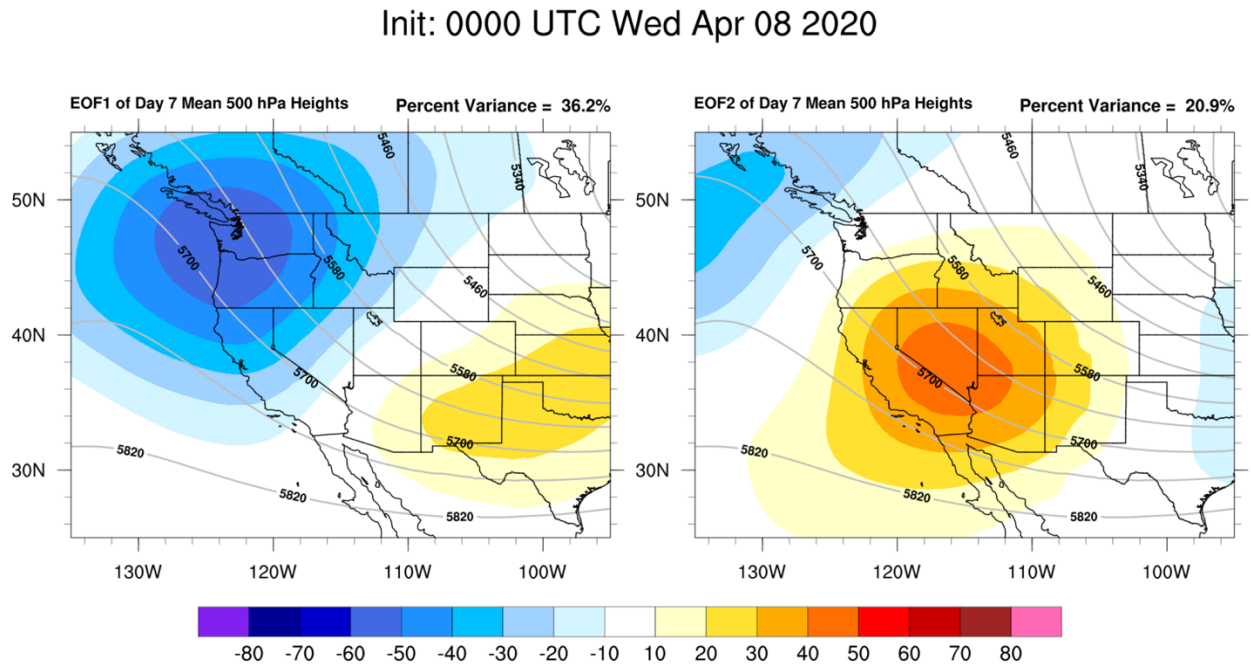
2. Getting Oriented to the WPC Cluster Analysis Tools

Our orientation will be based primarily on the 3-7 Day Clusters. We will walk through each of the options available on the WPC Cluster Analysis website. Please follow the link below: https://origin.wpc.ncep.noaa.gov/hmt/wk2/day_3_7/view.php

2a. 500-mb EOF Patterns

Quick summary: The two leading EOFs for 500-mb heights (i.e., the two patterns explaining the spread amongst the ensemble members). The first EOF always explains the greatest percentage of spread in the 500-mb height field; the second EOF explains the second greatest percentage of spread.

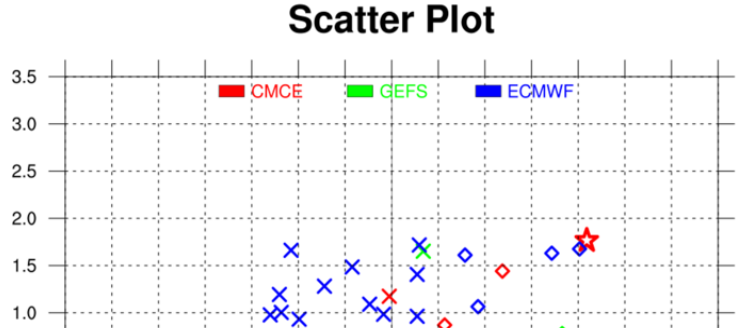
Select the options "Day 7", "500 mb EOF Patterns", and "West" from the toolbar above the graphics. Shown are the 2 leading EOFs based on uncertainty amongst the ensemble members for 500-mb heights. More details on EOFs are provided in the section, "Understanding Empirical Orthogonal Functions (EOFs) and Principal Components (PCs)". An example from 0000 UTC Wed Apr 08 is shown below:



Note: A **dipole** centered on the ensemble-mean location of the ridge or trough typically **indicates position and/or timing uncertainty in the location of the ridge or trough**. A **monopole** centered near the ensemble-mean location of the ridge or trough typically **indicates uncertainty in the amplitude of the ridge or trough**. The sign (and color) of the dipole does not really matter. It just helps us identify which members look similar (positive) or dissimilar (negative) to that EOF.

2b. Cluster Phase Space

Quick summary: A scatter plot



showing which cluster each individual ensemble member falls into.

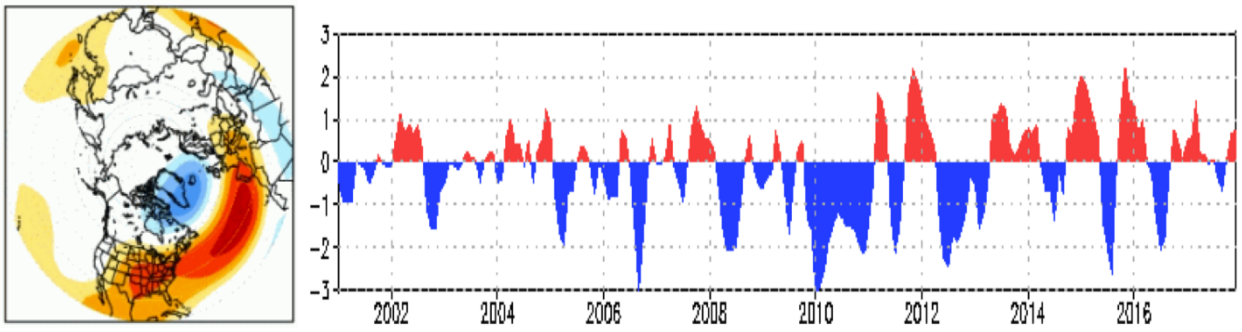
Select the options “Day 7”, “Cluster Phase Space”, and “West” from the toolbar above the graphics. Shown is a phase space diagram with the two principal components (PCs) of each ensemble member plotted. Colors indicate the ensemble system, and symbols indicate both the cluster of each member as well as the ensemble mean and deterministic run. More details on PCs are provided in the section below, “ Empirical Orthogonal Functions (EOF) and Principal Components (PCs)”. An example from 0000 UTC Wed Apr 08 is shown to the right:

2c. Empirical Orthogonal Functions (EOFs) and Principal Components (PCs)

While not an actual page on the WPC Cluster Analysis website, this topic is critically important to understanding the EOFs and Phase Space diagram and hence, it is included here.

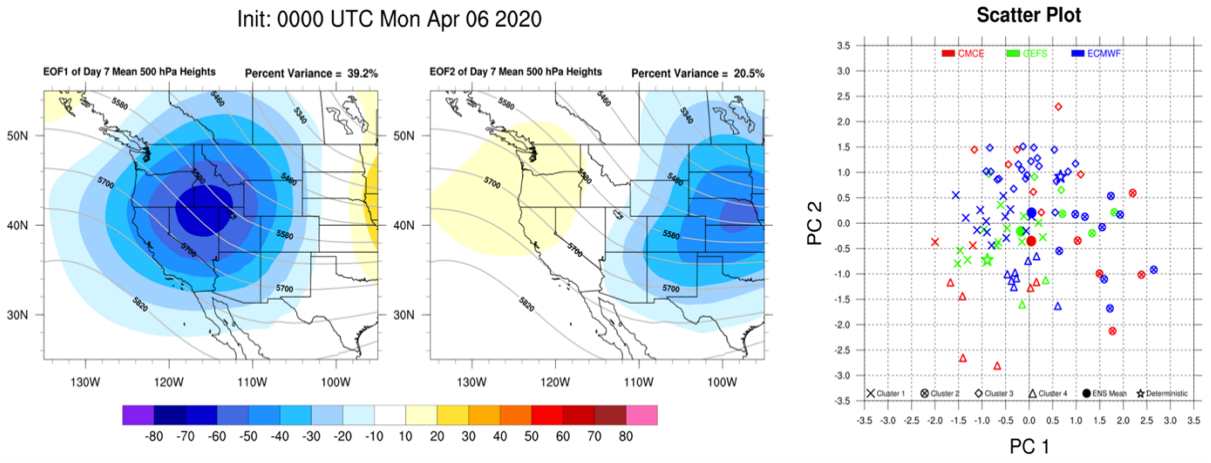
In climate studies, EOF analysis is often used to study possible spatial modes (i.e., patterns) of variability and how they change with time (e.g., the North Atlantic Oscillation [NAO]) (ref).

For example, consider the NAO Index. If you examine a multi-decadal time series of geopotential height anomalies over the North Atlantic Ocean, you can statistically determine the leading spatial modes (or patterns) of variability over this domain. These modes are called EOFs, and the mode that explains the greatest amount of variability, often called the leading mode or EOF 1, is shown below at left. Going back to our time series, we then calculate how well the spatial pattern observed at each time step matches the EOF 1. Spatial patterns that are more similar are increasingly positive numbers and patterns that are more dissimilar are increasingly negative numbers. We can then plot this time series, sometimes called PC 1, and the result is the NAO time series shown below at right.



Now, consider the WPC Clusters. Our domain has changed from the North Atlantic Ocean to the Western United States. Instead of just one leading EOF (EOF 1), the two leading EOFs (EOF1 and EOF2) are calculated (shown below, at left). Instead of EOFs calculated based on variability in geopotential height over time steps, EOFs are now calculated based on variability in geopotential height over ensemble members. If a particular ensemble member looks similar to the EOF, then it will have a positive PC. If it looks dissimilar, it will have a negative PC. So for EOF1 and EOF2, we could do as above, plotting the ensemble member on the x-axis and the

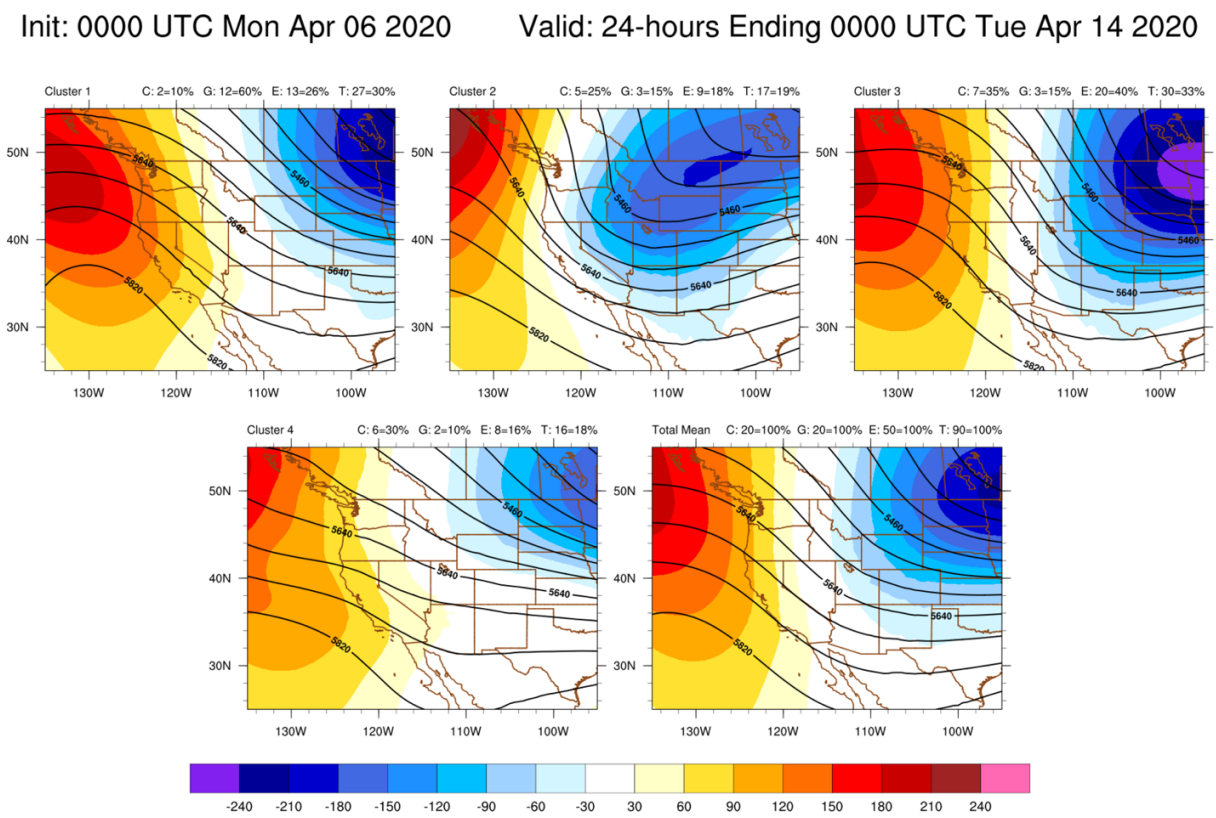
corresponding PC (degree of similarity/difference between that member and the EOF) on the y-axis. Instead, we plot them both onto the phase space diagram (shown below, at right). This allows us to more clearly see how individual ensemble members are distributed according to their PCs (which, again, represent how closely they match the EOFs).



Here's another way to think through this process: The goal of ensemble-based cluster analysis is to group individual ensemble members into clusters such that the total spread within clusters, summed over all clusters, is minimized. There is, of course, still spread within each cluster. Think of it as packing widgets (ensemble members) into boxes (clusters) and trying to select boxes such that you have the least amount of space left over. In this case, there are 90 widgets and you want 4 boxes. Cluster analysis optimizes those box sizes for you. In other words, 4 boxes (clusters) have been chosen such that all 90 widgets (members) fit into them with the least possible amount of space left over.

The key advantage of cluster analysis is that you can see what kinds of similar solutions the ensemble members are clustered around. Returning to our analogy for the case shown below, cluster #3 represents the third box. There are 30 widgets in it (~33% of the total widgets), and the pattern shown depicts the average shape of those 30 widgets.

The deterministic runs of the GFS, EC, and Canadian models will also resemble one of the four "boxes" as well. Details in that higher-resolution solution may be useful when considering how to message the scenario it fits best in. However, realize that the next run of that deterministic model may change and look like a different "box". That's okay. If you were able to look at Ensemble member #4 from the GEFS, it would also "bounce around" from scenario to scenario, as would all the members. That's how an ensemble system works. The clustering approach allows us to not get focused on the "bounce" or "flip-flop" and instead extract useful information on how the total ensemble system is describing the uncertainty of the pattern. The main point is that we can leverage them all to message various worst (or best) case scenarios.



Note: It's important to point out that because the algorithm always selects 4 clusters, there will be times when it is "slicing the onion *really* thin" (i.e., times when the ensemble members are all in fairly good agreement). For this reason, one should always look at the EOFs to determine how much uncertainty actually exists, before diving into the clusters.

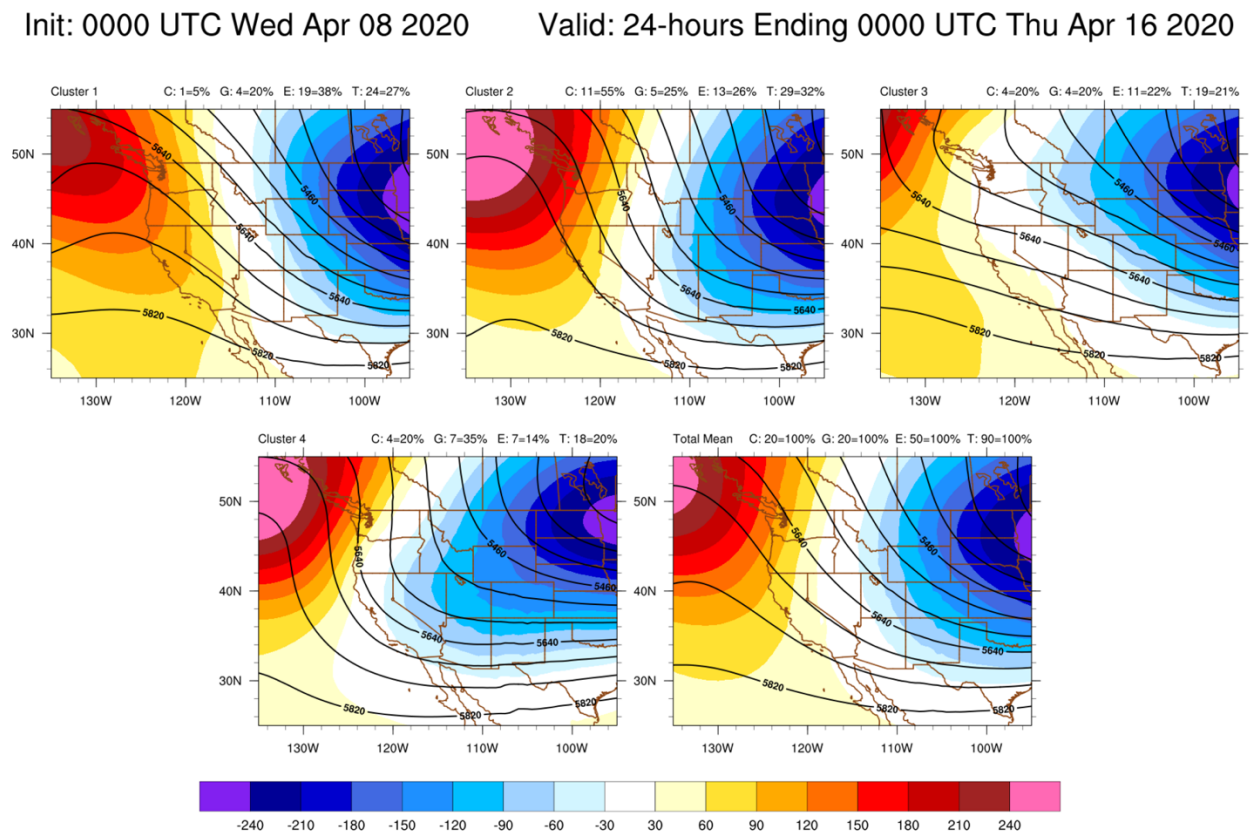
2d. 500-mb Height Clusters

Quick summary: The cluster-mean 500-mb heights and anomalies relative to climatology for each of the four clusters and the all-member mean.

Select the options "Day 7", "Cluster 500 Heights", and "West" from the toolbar above the graphics. Shown are the 4 ensemble member clusters (while any number of clusters can be used, WPC finds that 4 works best) and in the bottom right, the all-member mean. For each cluster, the black contoured lines show the cluster mean (in decameters), and the color shading shows the difference (in meters) between the cluster mean and the climatological mean (1980-2010 CFSR). At the top of each cluster is indicated the number and percentage of members from the CMCE (C), the GEFS (G), the EPS (E), and the total (T). An example from 0000 UTC Wed Apr 08 is shown below:

Caution: At times, one ensemble system will go "all in", with 90-100% of its members in one cluster. This often implies that the ensemble system in question is underdispersive (i.e., it does not accurately reflect the true spread), particularly when the other clusters are significantly

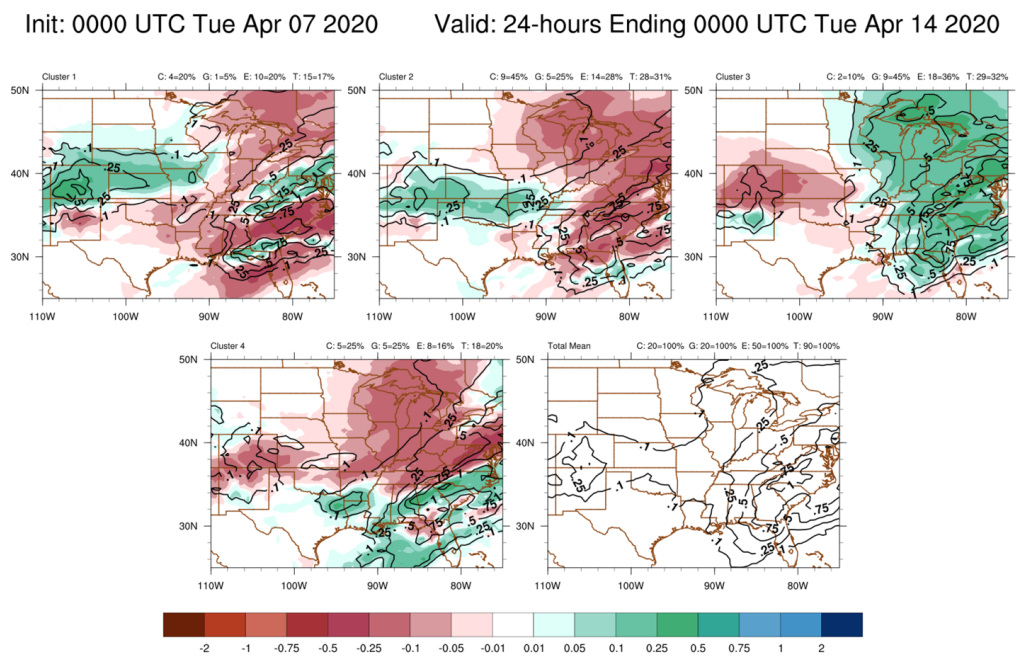
different. Often, the atmosphere will verify outside the reduced spread of such underdispersive ensemble output.



2e. QPF Clusters (currently 24-h QPF for Days 3-7 and 72-h QPF for Days 8-10)

Quick summary: The cluster-mean QPF and anomalies relative to the all-member mean for each of the four clusters.

Select the options “Day 7”, “Cluster 24-h QPF”, and “West” from the toolbar above the graphics. Shown are the 4 ensemble member clusters and in the bottom right, the all-member mean. For each cluster, the black contoured lines show the cluster mean (in inches), and the color shading shows the difference (in inches) between the cluster mean and the all-member mean (***not* the climatological mean, as is shown for 500-mb heights**). At the top of each cluster is indicated the number and percentage of members from the CMCE (C), the GEFS (G), the EPS (E), and the total (T). An example from 0000 UTC Wed Apr 07 is shown below:



Understanding the Color Scales for QPF and Max/Min Temperatures

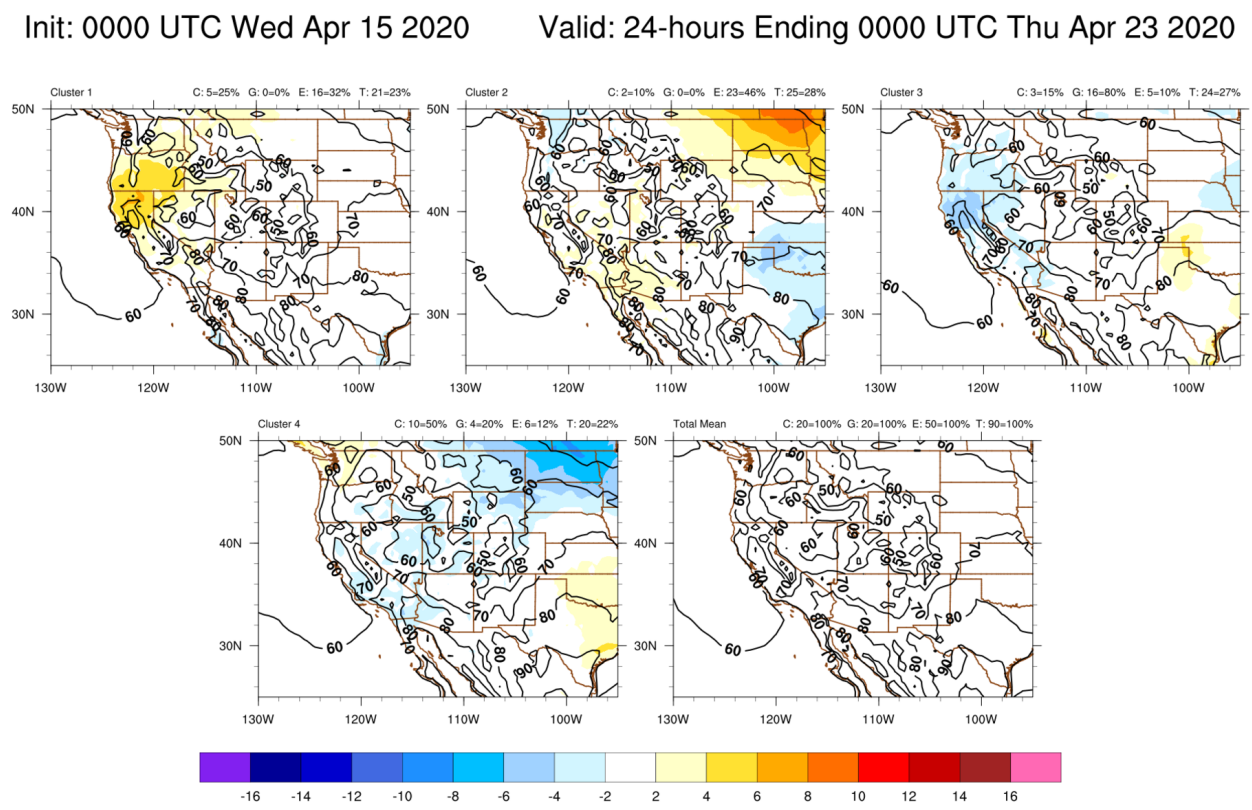
The clustered forecasts for QPF (and Max/Min Temperatures) contain a wealth of information, but to clarify their interpretation, let's focus on the example shown above, for QPF. Here, black contours show the cluster-mean QPF and the green/brown color shading shows the anomaly of cluster-mean QPF relative to the all-member mean QPF. *This is done to highlight how each cluster mean differs from the all-member mean*. The all-member mean QPF is shown in the bottom right, and since the anomaly of that all-member mean QPF relative to itself is zero everywhere, there is no color shading.

So remember, *the color shading is relative to the all-member mean, rather than climatology.* **This means that even in brown-shaded areas, precipitation may still be possible, and it may be a non-trivial, even impactful, amount**.

2f. Maximum/Minimum Temperature Clusters

Quick summary: The cluster-mean max/min temperature and anomalies relative to the all-member mean for each of the four clusters.

Select the options "Day 7", "Cluster Maximum/Minimum Temperatures", and "West" from the toolbar above the graphics. Shown are the 4 ensemble member clusters and in the bottom right, the all-member mean. For each cluster, the black contoured lines show the cluster mean (in F), and the color shading shows the difference (in F) between the cluster mean and the all-member mean (***not* the climatological mean, as is shown for 500-mb heights**). At the top of each cluster is indicated the number and percentage of members from the CMCE (C), the GEFS (G), the EPS (E), and the total (T). An example from 0000 UTC Wed Apr 15 is shown below:



Note: The clusters are based on 500-mb heights, so mesoscale features appearing in the clusters for QPF and Max/Min Temperature may be coincidental (i.e., not necessarily driven by the 500-mb pattern or processes associated with it). In addition, since there is no calibration, features associated with ensemble system resolution and biases can at times be seen, particularly when one ensemble system is going "all in" on a particular cluster.

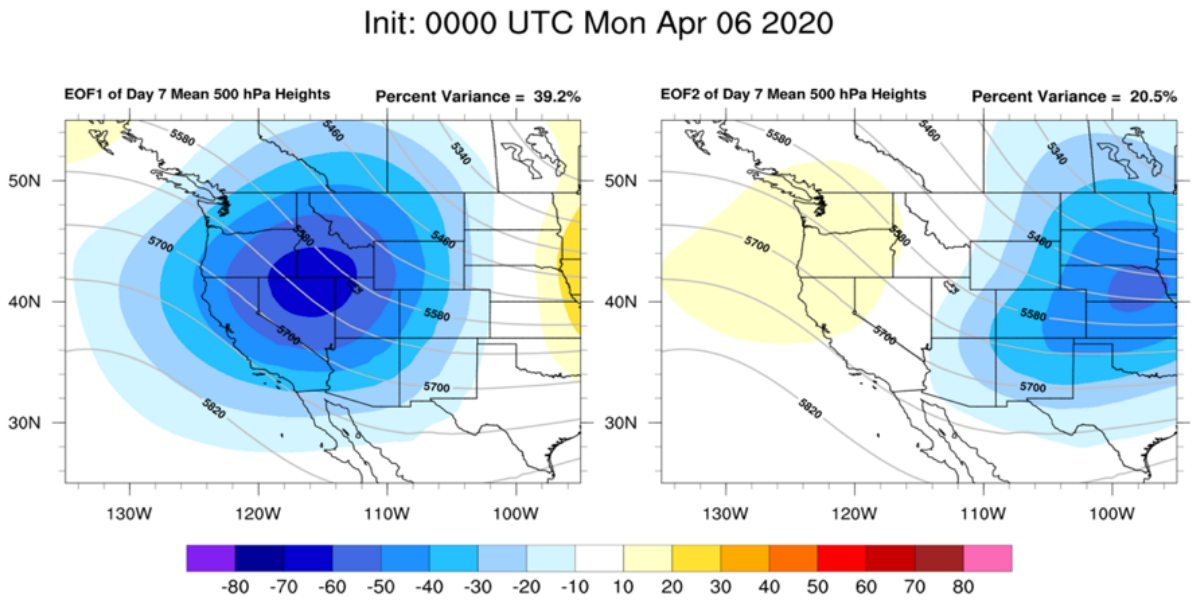
3. Additional Topics

3a. Is the EOF Positive or Negative? Understanding the Sign

One of the most confusing aspects of EOFs and PCs are the resulting signs (i.e., positive or negative). As meteorologists, we immediately think: "*warm colors = ridging*", "*cold colors = troughing*". But that's not how the EOFs and PCs should be interpreted. **It's important to understand that the sign results purely from a statistical calculation and has no physical meaning. It only takes on meaning once you begin to interpret it.**

For example, in EOF1 shown in Section 3 (and below), the gray contours show the ensemble-mean 500-mb height pattern over WR. The color shading indicates the ensemble members' spread for 500-mb heights. The large values (< -60) over WR indicate large spread amongst the ensemble members' 500-mb heights in this region—many show more ridging than the ensemble-mean; many show more troughing than the ensemble mean. The negative sign

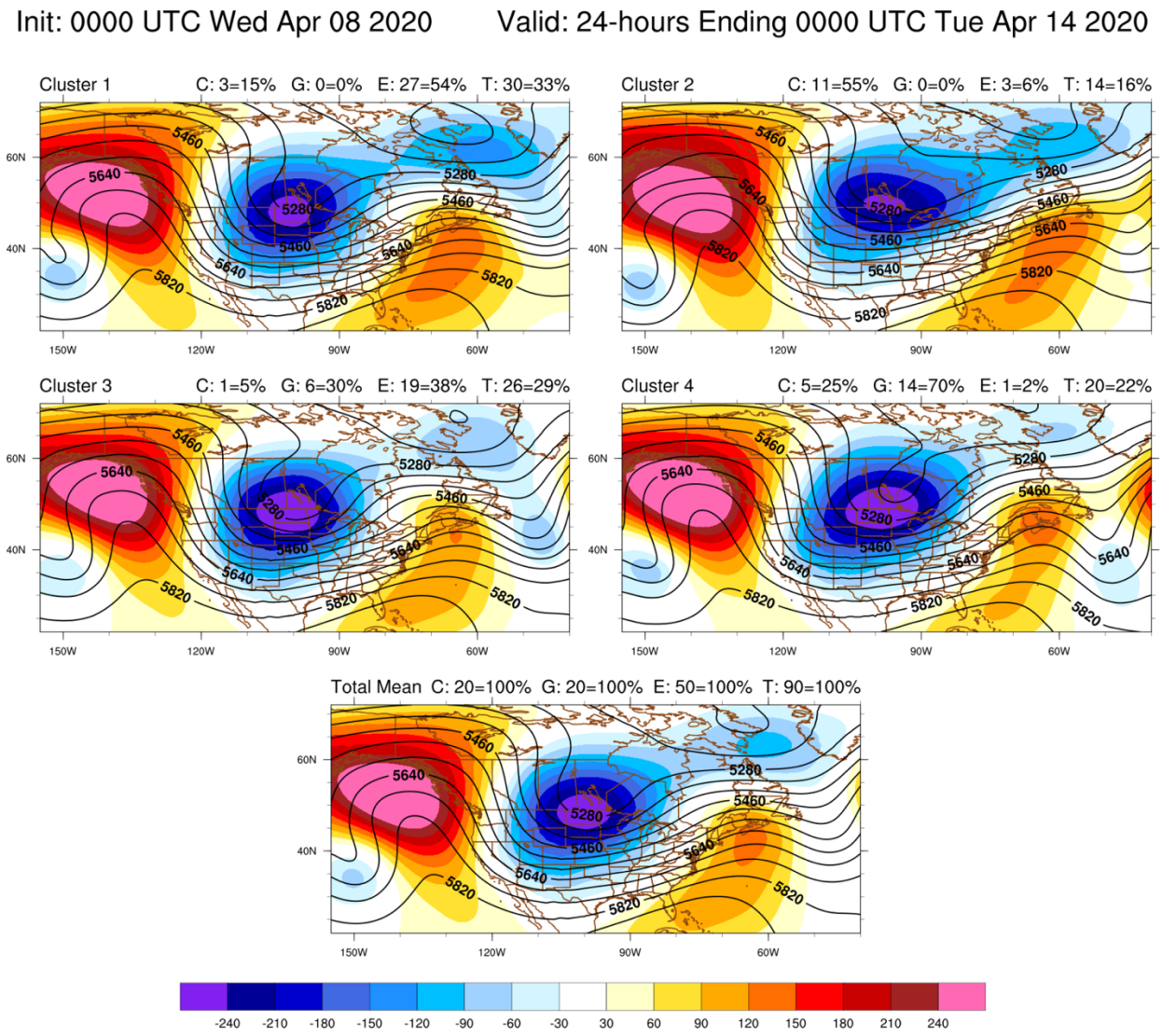
(blue shading) doesn't say anything about troughing or a preference for it in this region. It merely indicates that those members showing more troughing than the ensemble mean in this area will have a PC1 that is *positive* for EOF1, because this pattern is *similar* to EOF1. Likewise, those members showing more ridging than the ensemble mean in this area will have a PC1 that is *negative* for EOF1, because this pattern is *dissimilar* to EOF1.



3b. How the Chosen Domain Affects the Clusters

The domain chosen has a major influence on the resulting clusters. As domain size increases, it becomes increasingly likely that upper-level features producing the greatest uncertainty amongst ensemble members are far away from the forecast region of interest. Since the region of greatest uncertainty will dominate the EOFs and the resulting clusters, important information in the region of forecast interest can be lost.

For example, consider the forecast for Detroit, MI (DTW; located in southeastern MI) based on the CONUS domain shown below. The 500-mb heights over DTW (~552 dm) are incredibly consistent across clusters. However, these clusters are driven not just by the extent of troughing over the northern Great Plains, but by the extent of ridging over eastern North America, and the extent of ridging over the Gulf of Alaska. There are even features over the central Atlantic playing a role in the clustering! Therefore, we might be missing key variance in the vicinity of DTW.



Now let's zoom in and look at the CENTRAL domain, shown below. The highest 500-mb heights over DTW are ~556 dm in Cluster #3, and the lowest 500-mb heights over DTW are ~544 dm in Cluster #4. Because our EOFs and resulting clusters are no longer based on uncertainties associated with far away features, we've obtained clusters of both higher- and lower-height members than we were able to before, and these clusters show more of (though not all of) the spread amongst ensemble members. This provides much more information for constructing potential forecast scenarios.

Init: 0000 UTC Wed Apr 08 2020 Valid: 24-hours Ending 0000 UTC Tue Apr 14 2020