Rutz Jonathan, J. (Orcid ID: 0000-0003-4337-0071)

Shields Christine, A (Orcid ID: 0000-0003-4481-1377)

Guan Bin (Orcid ID: 0000-0002-8428-5984)

Ullrich Paul (Orcid ID: 0000-0003-4118-4590)

O'Brien Travis, Allen (Orcid ID: 0000-0002-6643-1175)

Leung L. Ruby (Orcid ID: 0000-0002-3221-9467)

Wehner Michael, F (Orcid ID: 0000-0001-5991-0082)

Brands Swen (Orcid ID: 0000-0002-3254-0277)

Collow Allison, B. (Orcid ID: 0000-0002-3566-3889)

Goldenson Naomi (Orcid ID: 0000-0003-0656-4168)

Kashinath Karthik (Orcid ID: 0000-0002-9311-5215)

Krishnan Harinarayan (Orcid ID: 0000-0001-8018-0547)

Lavers David (Orcid ID: 0000-0002-7947-3737)

Magnusdottir Gudrun (Orcid ID: 0000-0001-6079-5886)

Prabhat Mr (Orcid ID: 0000-0003-3281-5186)

Qian Yun (Orcid ID: 0000-0003-4821-1934)

Ramos Alexandre, M. (Orcid ID: 0000-0003-3129-7233)

Sellars Scott, Lee (Orcid ID: 0000-0003-0778-8964)

Shulgina Tamara (Orcid ID: 0000-0002-7381-3331)

Tomé Ricardo (Orcid ID: 0000-0003-0657-7057)

Wilson Anna, Maria (Orcid ID: 0000-0001-7342-1955)

Viale Maximiliano (Orcid ID: 0000-0002-4093-4129)

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1029/2019JD030936

The Atmospheric River Tracking Method Intercomparison Project (ARTMIP): Quantifying Uncertainties in Atmospheric River Climatology

Jonathan J. Rutz and Co-Authors (Next Page)

Science and Technology Infusion Division, National Weather Service, Salt Lake City, UT

Submitted to Journal of Geophysical Research: Atmospheres
Revised Draft Manuscript
20 November, 2019

Authors: Jonathan J. Rutz[1], Christine A. Shields[2], Juan M. Lora[3], Ashley E. Payne[4], Bin Guan[5], Paul Ullrich[6], Travis O'Brien[7], L. Ruby Leung[8], F. Martin Ralph[9], Michael Wehner[10], Swen Brands[11], Allison Collow[12], Naomi Goldenson[13], Irina Gorodetskaya[14], Helen Griffith[15], Karthik Kashinath[16], Brian Kawzenuk[9], Harinarayan Krishnan[10], Vitaliy Kurlin[17], David Lavers[18], Gudrun Magnusdottir[19], Kelly Mahoney[20], Elizabeth McClenny[6], Grzegorz Muszynski[16,17], Phu Dinh Nguyen[21], Mr. Prabhat[16], Yun Qian[8], Alexandre M. Ramos[22], Chandan Sarangi[8], Scott Sellars[23], T. Shulgina[9], Ricardo Tome[22], Duane Waliser[5], Daniel Walton[24], Gary Wick[19], Anna M. Wilson[9], Maximiliano Viale[25]

Affiliations:

1. Science and Technology Infusion Division, National Weather Service Western Region Headquarters, National Oceanic and Atmospheric Administration, Salt Lake City, UT 84138, USA
2. Climate and Global Dynamics Division, National Center for Atmospheric Research, Boulder, CO 80302, USA
3. Department of Geology and Geophysics, Yale University, New Haven, CT 06511, USA
4. Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI 48109, USA
5. Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA
6. Department of Land, Air and Water Resources, University of California, Davis, CA 95616, USA
7. Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
8. Earth Systems Analysis and Modeling, Pacific Northwest National Laboratory, Richland, WA 99354, USA
9. Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, La Jolla, CA 92093, USA
10. Computational Chemistry, Materials, and Climate Group, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
11. MeteoGalicia – Xunta de Galicia, Santiago de Compostela, Spain
12. Universities Space Research Association, Columbia, MD 21046, USA
13. Center for Climate Science, University of California Los Angeles, Los Angeles, CA 90095, USA
14. Centre for Environmental and Marine Studies, University of Aveiro, 3810-193 Aveiro, Portugal
15. University of Reading, Reading, United Kingdom
16. Data & Analytics Services, National Energy Research Scientific Computing Center (NERSC), Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
17. Department of Computer Science, University of Liverpool, Liverpool, UK
18. European Centre for Medium-Range Weather Forecasts, Reading, RG2 9AX, UK
19. Department of Earth System Science, University of California Irvine, Irvine, CA 92697, USA
20. Physical Sciences Division, Earth System Research Laboratory, National Oceanic and Atmospheric Administration, Boulder, CO 80305, USA
21. Department of Civil & Environmental Engineering, University of California Irvine, Irvine, CA 92697, USA
22. Instituto Dom Luiz, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisbon, Portugal
23. National Science Foundation, Office of Advanced Cyberinfrastructure, Alexandria, VA 22314, USA
24. Institute of the Environment and Sustainability, University of California, Los Angeles, CA 90095, USA
25. Instituto Argentino de Nivologia, Glaciologia y Ciencias Ambientales, CCT-CONICET, Mendoza, Argentina

Abstract

Atmospheric rivers (ARs) are now widely known for their association with high-impact weather events and long-term water supply in many regions. Researchers within the scientific community have developed numerous methods to identify and track of ARs—a necessary step for analyses on gridded data sets, and objective attribution of impacts to ARs. These different methods have been developed to answer specific research questions, and hence use different criteria (e.g., geometry, threshold values of key variables, time dependence). Furthermore, these methods are often employed using different reanalysis data sets, time periods, and regions of interest. The goal of the Atmospheric River Tracking Method Intercomparison Project (ARTMIP) is to understand and quantify uncertainties in AR science that arise due to differences in these methods

This paper presents results for key AR-related metrics based on 20+ different AR identification and tracking methods applied to MERRA v2 reanalysis data from January 1980 through June 2017. We show that AR frequency, duration, and seasonality exhibit a wide range of results, while the meridional distribution of these metrics along selected coastal (but not interior) transects are quite similar across methods. Furthermore, methods are grouped into criteria-based clusters, within which the range of results is reduced. AR case studies and an evaluation of individual method deviation from an all-method mean highlight advantages/disadvantages of certain approaches. For example, methods with less (more) restrictive criteria identify more (less) ARs and AR-related impacts. Finally, this paper concludes with a discussion and recommendations for those conducting AR-related research to consider.

1. Introduction

Over the past several years, interest in atmospheric river (AR) science and applications has increased rapidly. Beyond the now well-known impacts of heavy rain and flooding (e.g., Neiman et al. 2008; Ralph et al. 2013; Lamjiri et al. 2017), ARs have been shown to have applications in areas as diverse as avalanche hazard (Hatchett et al. 2017), dust transport (Ault et al. 2011), and post-fire debris flows (Oakley et al. 2017). Furthermore, the study of ARs has become global in scope, and international in terms of participation, as evidenced by the well-attended 2018 International Atmospheric Rivers Conference (Ramos et al. 2019). The American Meteorological Society (AMS) Glossary of Meteorology defines an AR as "a long, narrow, and transient corridor of strong horizontal water vapor transport that is typically associated with a low-level jet stream ahead of the cold front of an extratropical cyclone". The development of this definition, a process described by Ralph et al. (2018a), was marked by open engagement with the atmospheric and geosciences community throughout the process, and should be considered a major success in the field. However, the elegance of this definition depends on its *qualitative* description of ARs, whereas, in practice, the peer-reviewed literature contains dozens of *quantitative* definitions of ARs, as needed in analysis and modeling. These quantitative definitions are manifested as different AR identification and tracking methods that researchers have developed to answer a wide variety of questions. Note also that the large majority of these methods were developed prior to the development of the AR definition within the AMS Glossary of Meteorology.

Each individual method identifies and/or tracks ARs on the basis of selected criteria being met, as summarized in Fig. 1. A first step in development of these methods is often the choice of a thresholding variable and magnitude, which serves as the minimum requirement for identifying ARs. The thresholding variable can be integrated water vapor (IWV; e.g., Wick et al. 2013), but is most commonly IWV transport (IVT), and the magnitude can be either *absolute* (e.g., IVT $\geq$ 250 kg m^{-1} s^{-1}; e.g., Rutz et al. 2014) or *relative* (e.g., IVT $\geq$ 85th percentile of local climatological IVT; e.g. Lavers et al. 2012). Research has shown that using IVT extends medium-range predictability for high-impact hydrological events (Lavers et al. 2017), and recent field campaigns have used probabilistic IVT forecasts to determine AR location and intensity (Cordeira et al. 2017). Once the thresholding process is applied to the data, features meeting or exceeding the threshold are examined with respect to geometric parameters such as length, width, shape, axis, and orientation. Throughout this study, methods with lower-magnitude thresholds and less geometric requirements will generally be referred to as "less restrictive methods," whereas methods with higher-magnitude thresholds and more geometric requirements will generally be referred to as "more restrictive methods." Note also that some methods, particularly those based on machine learning techniques, do not directly use any thresholds as requirements. Temporal requirements may also be chosen (i.e., either AR identification is independent of time [*time slicing*], or it is dependent on criteria being met for a certain duration [*time stitching*]). The choices described above lead to many possible permutations, and while some methods feature similar criteria, others vary widely. Of course, in addition to using different identification and tracking methods, many researchers examine different regions, using different data sets, and different periods of

records, to do so. More recently, machine learning techniques have been developed to identify and track ARs (e.g., Radic et al. [2015]; Mudigonda et al. [2017]; Muszynski et al. [2019]).

These different methods produce differences in AR climatologies and, consequently, differences in the impacts attributable to ARs. These differences produce uncertainty in operational weather research and forecasting, water management, and climate projections, which require a current baseline of AR climatology and impacts to assess future changes. The differences in identified ARs that can be observed during a single event are highlighted using a case from 0000 UTC 15 February 2014, shown in Fig. 2. Notice that some methods identify an AR only over the greatest values of IVT offshore, others extend near just inland of the coast, and some extend well into the continental interior. These differences have major consequences. For example, one question the water management community might ask is, "what fraction of precipitation is attributable to ARs, and how might that change under future climate change scenarios?" Before even exploring climate change scenarios, one needs to answer the first question, and the answer depends on which method is chosen. Fig. 3 shows the fraction of cool-season or annual precipitation attributable to ARs based on three studies (Dettinger et al. [2011]; Rutz et al. [2014]; Guan and Waliser [2015]). These studies found broadly similar spatial patterns, but quite widely varying values from southern California (~15–35%) to coastal Washington (~25–60%). It is worth noting that in addition to different AR identification methods, these studies also used different data sets, different periods of record, and different methods of attributing precipitation to ARs, all of which contribute to this range.

It is critical to remember that each AR identification and tracking method was developed to answer a specific question or set of questions, and that these questions vary widely from one study to the next. Having a sense of these original questions better informs the reader as to the original intent or goal of each method, as described in the supplemental material provided by method developers. For example, Ramos et al. (2015) examined the relationship between persistent ARs and extreme precipitation over the Iberian Peninsula; Rutz et al. (2014) identified ARs and their impacts over the complex topography of the western United States; and Guan and Waliser (2015) produced a global climatology of ARs and their characteristics. Furthermore, Shields and Kiehl (2016) and Gershunov et al. (2017) explored the climate scale variability of ARs along the North American West Coast. Still other methods are using machine learning techniques to determine whether ARs can be identified without the use of defined thresholds (e.g., Muszynski et al. [2019]). With such a variety of different questions asked, and such different goals pursued, it should not be surprising that many different results have been found. Nevertheless, a growing awareness of the uncertainties that these differences produce has led to the development of a community-based project to better understand and quantify them.

The goal of the Atmospheric River Tracking Method Intercomparison Project (ARTMIP; Shields et al. 2018) is to quantify and understand the uncertainties in AR climatology (e.g., frequency, duration, and intensity), precipitation, and related impacts that arise from different AR identification and tracking methods, and how uncertainties in these AR-related metrics

may change in the future. Furthermore, ARTMIP aims to understand the implications of those uncertainties in terms of our recent, current, and future climate. A few recent studies have focused on this topic. Huning et al. (2017) examined the sensitivity of AR-attributable snowfall in California's Sierra Nevada to AR detection methods based on two different AR catalogs. Guan and Waliser (2015) examined the sensitivity of AR detection to intensity/geometry thresholds and input datasets, but only based on a single AR detection algorithm. Ralph et al. (2018b), in an initial pre-ARTMIP study, quantified uncertainties in AR-related metrics using ~10 AR detection algorithms but focused on only one location along the California coast. This paper provides a systematic and global inter-comparison between different methods by quantifying the uncertainty in current (1980–2017) AR climatology on a global scale, using over 20 AR identification and tracking methods. To do so, it leverages a variety of metrics, which are described in more detail in the following sections. An assessment of method-related uncertainty affecting AR climatology under climate change scenarios will be the subject of another paper, discussed at the end of Section 4.

2. Data and Methods

The progression of ARTMIP is organized into "tiers," and this study is a summary of results from the Tier 1 phase of the project. The data used in Tier 1 of ARTMIP are described at length in Shields et al. (2018), and a brief overview is also given here.

A key aspect of ARTMIP is that analyses are performed using the same atmospheric dataset, over the same period of record, and over the entire globe. This enables a clean comparison of AR-related metrics across all methods, whereas previous studies used different atmospheric data sets, different periods of records, and examined only certain regions. Note, however, that some methods' criteria explicitly limit their results to certain regions, and a mask is used to indicate these regions. Basic quantities such as IWV and IVT, which is often a derived variable, were pre-computed for ARTMIP to ensure that all algorithms use exactly the same data. The atmospheric data for these calculations comes from the MERRA-2 reanalysis (Gelaro et al., 2017) for the period of January 1980 through June 2017, at a horizontal resolution of 0.625×0.5° and a 3-h temporal resolution. The ARTMIP catalogs are then produced by developers applying their identification and tracking methods to these data. For each 3-h time slice, each grid point is flagged with a 0 for "AR conditions do not exist" or a 1 for "AR conditions exist". Catalogues produced for Tier 1 as well as the source MERRA-2 data used by all ARTMIP participants are available on the Climate Data Gateway (CDG). MERRA-2 source data can be found at https://doi.org/10.5065/D62R3QFS (NCAR/UCAR Climate Data Gateway, 2018), and ARTMIP Tier 1 output data catalogues, also housed on the CDG, at doi:10.5065/D6R78D1M. Table 1 summarizes all the methods participating in ARTMIP with notation specifying Tier 1 algorithms only.

Key results are presented along selected, roughly meridional transects along the North American West Coast, through interior western North America, and along the European West Coast (Fig. 4). These transects are selected because most regional methods have been developed, and produce data, for one of these two regions. The coastal transect points are

determined by selecting all MERRA-2 reanalysis grid points that have fractional land/sea cover between 32–55°N (and 130–115°W) for North America, and between 35–62°N (and 15°W –10°E) for Europe. The interior transect points are determined by subjectively selecting grid points that represent a significant topographic "crest" between 32–55°N. The interior transect facilitates comparison between results for AR-related metrics along a coastline, which lies downstream of an ocean, and results for AR-related metrics over an interior region, which lies downstream of, and embedded within, complex topography.

This paper will present a number of results based on grouping methods into "clusters" that have similar approaches to identifying ARs. Throughout this section, refer to Fig. 5 for a summary of which clusters each method is grouped into, and Table 1 for more in-depth information regarding each method. Note also that many groups have joined ARTMIP and contributed data since the beginning of this analysis and are not listed here, but can be found online. Their areas of focus include South America (Viale et al. 2018) and Polar regions (Gorodetskaya et al. 2014), among others. The first key cluster pair is that differentiating between methods using *absolute* thresholds (e.g., IVT $\geq$ 250 kg m^{-1} s^{-1}) and methods using *relative* thresholds (e.g., IVT $\geq$ 85th percentile of climatological IVT). This is done because these are fundamentally different ways of identifying and tracking ARs, and the visualization of results benefits from the distinction. Throughout this paper, the terms *absolute* and *relative* will be italicized when used in this context. There is also a subtle, but important difference among the *relative* methods themselves: those whose thresholds vary as a function of latitude (*latitude-dependent relative* methods) and those that do not (*latitude-independent relative* methods). *Latitude-dependent relative* methods use thresholds based on the climatology of each grid point, and can be expected to produce smaller meridional gradients in AR statistics. *Latitude-independent relative* methods use one threshold based on the climatology of a given region, and can be expected to produce larger gradients in AR statistics, which will likely be more similar to results produced by *absolute* methods. Furthermore, this paper includes one method based on machine learning (TDA_ML; Muszynski et al. 2019), which defies many of the threshold-based groupings outlined above. It is currently employed over the western U.S., but could be readily applied to other regions.

Another distinction made in this study, and key cluster pair, is that between *global* and *regional* methods, which simply describes the area over which the method was originally developed and applied. Masks for each regional method are found in the supplemental material of the experimental design paper, Shields et al. (2018).

Finally, a subjective distinction will at times be made between methods that are either less restrictive or more restrictive. Here, "less restrictive" generally denotes a method or methods with less restrictive criteria required for AR identification, leading to a greater number of ARs being identified. Similarly, "more restrictive" generally denotes a method or methods with more restrictive criteria required for AR identification, leading to a smaller number of ARs being identified. Future work will quantify the "restrictiveness" of such methods, but these generalizations will be used throughout this paper.

3. Results

This section describes climatological characteristics of ARs based on the ARTMIP methods used in Tier 1. These characteristics are highlighted via a few key metrics, including AR frequency and duration. Here, results are presented either along selected transects or in a zonal-mean framework to facilitate a more focused analysis.

3.1 AR Frequency

This section discusses AR frequency, which is defined as the percentage of time that a given location is experiencing AR conditions (i.e., is located within the spatial footprint of an AR). For example, if a given method produces an AR frequency of 10% at some location, it means that this location is within the spatial footprint of ARs, as identified by that method, 10% of the time from January 1980 through June 2017, inclusive of all months. AR frequency along the North American and European West Coasts, as well as through interior western North America, varies greatly as a function of method used (Fig. 6). Focusing on the North American West Coast, nearly all methods exhibit a rapid increase in AR frequency from a minimum near 32°N towards a maximum near 45°N, followed by a more gradual decrease northward toward 56°N (Fig 6, top). This distribution closely resembles that of North Pacific storm track density shown by Lukens et al. (2018; their Fig. 4b), among others. In general, both less restrictive criteria and *absolute* thresholds lead to more dramatic changes in AR frequency as a function of latitude, with the "Rutz" method exhibiting the greatest maximum (~14%) and range (~12%) along this transect. The AR frequency of the "Guan_Waliser" method is an exception to the generalized statements above—it exhibits a gradual increase as a function of latitude and a small range (~4%) relative to the number of events it identifies. A similar behavior is also seen with Brands_v1. These characteristics arise from the fact that they are both percentile-based *relative* methods that use a *latitude-dependent* IVT threshold, where the direct influence on AR frequency from the climatological meridional gradient in IVT tends to be smoothed out. Such a smoothing effect is less obvious with "Brands_v2" and "Brands_v3", likely because the fixed lower limit of IVT becomes dominant compared to the less restrictive percentile thresholds in these two methods, making them inclined toward *absolute* methods. In a similar sense, *relative* methods that use *latitude-independent* IVT thresholds ("Payne", "Lora_NPac", "Lora_global") agree better with *absolute* methods, because for a given region the IVT threshold in these latitude-independent *relative* methods is nothing but a fixed value. Relative methods can vary substantially in their methodology, from percentile- and climatology-based thresholds ("Brands", "Guan and Waliser", "Lavers", "Lora", "Mundhenk", "Payne and Magnusdottir", "Ramos", "Viale", "Walton") to thresholds based on spatial anomalies ("Gorodetskaya", "Shields and Kiehl"). Interestingly, the AR frequency from the machine learning method, "TDA_ML", is characterized by a maximum of just 2% near 39°N, and declines to 0% north of 45°N, where many methods produce their larger frequency values. Muszynski et al. (2019) note that this method frequently produces false negatives (i.e., it fails to detect ARs) when an AR merges with another AR or "some other event with high concentration of water vapor and similar topological structure, such as an extratropical cyclone." This would happen with higher frequency in the more active storm track at latitudes north of 45°N, which may explain the rapid drop-off in AR detection associated with this method.

Focusing on the selected transect through interior western North America, the AR frequency is greatly reduced for nearly all methods (Fig. 6, center). However, the "Guan_Waliser" method is a remarkable outlier here, as it exhibits an AR frequency (~8 – 10%) only slightly lower than at the same latitudes along the North American West Coast. This clearly results from being a less restrictive and *relative* method. The AR frequency of other methods is much lower (~1–4%), with the "Rutz" and "Brands_v1" methods, owing to their less restrictive thresholds, being the largest of these at most latitudes. Most other methods exhibit a coastal maximum near 45°N (albeit of varying magnitude) that shifts northward to an interior maximum near 48°N. This shift arises because ARs making landfall near 45°N preferentially extend inland towards the east-northeast along the relatively low-elevation corridor of the Columbia River Basin, as shown by Rutz et al. (2015, their Fig. 3). Additional, secondary, corridors of inland penetration are located south of 32°N and north of 52°N, and all of these corridors play an important role in heavy precipitation events (Alexander et al. 2015) and growth of vegetation (Albano et al. 2017) over interior regions. In contrast, areas downstream of major topographic barriers feature a larger decrease from coastal to interior frequency due to AR decay, as moisture is more effectively removed by orographic precipitation. This is particularly true for ARs making landfall between ~32–38°N, which are severely disrupted by the southern Sierra Nevada Mountains (elevation 2-4 km), drastically lowering the inland frequency between ~38–44°N (following a typical trajectory of inland penetration; see Rutz et al. 2015, their Fig. 3).

Along the European West Coast, "Rutz" identifies the greatest AR frequency nearly everywhere, and diverges markedly from "Guan_Waliser" between 44–60°N. This divergence is likely due to a higher climatological value of IVT at these latitudes, which generally causes *absolute* methods to identify a greater number of ARs than *relative* methods. Other methods such as "Lora", "Mundhenk", and "Tempest" follow a distribution very similar to "Rutz", but with smaller amplitudes. Once again, the fairly good agreement between the *absolute* method of Rutz and *latitude-independent* method of Lora is not surprising. The dramatic jump in AR frequency near 45°N may be due to some combination of climatology (i.e., placement of the storm track; e.g., Lukens et al. 2018) and the greater number of coastal transect points at latitudes north of 45°N. It is worth noting that during the ARTMIP one-month experiment described in Shields et al. (2018; their Fig. 3), the human-control analysis yielded a greater AR frequency than any automated method along both coastal transects. (The human-control analysis consisted of two graduate students counting "by eye" all ARs making landfall for the North American and European coastlines for the month of February 2017.)

Clearly, different AR identification and tracking methods produce widely varying results for AR frequency along coastal transects. It is important to remember that the criteria of each individual method used in ARTMIP have been developed to answer specific scientific questions, often driven by regional and/or impacts-specific considerations. Since different questions were asked, it should be no surprise that different methods are used and different results produced. Nevertheless, while the ARTMIP methods do not agree on absolute values of AR frequency, they do exhibit remarkable agreement in their latitudinal distribution,

except for an outlier ("Guan_Waliser"), which, among all the methods examined in this study, is the only *relative* method that both has a global coverage and uses percentile-based, latitude- and longitude-dependent IVT thresholds. Also, despite generally large inter-method differences, Brands_v2 and CONNECT500 yield practically identical results on both continents, which is quite surprising since the two methods have been developed independently.

The ARTMIP methods' good agreement regarding the latitudinal distribution of AR frequency is more clearly seen by normalizing each method as follows: for each method, the largest and smallest value along a given transect are given values of 1 and 0, respectively, and all values are then normalized to this scale (Fig. 7). For example, if the largest and smallest AR frequency along a given transect are 24% and 8%, respectively, a value of 12% will be normalized to 0.25. Exceptions to good agreement (e.g., "TDA_ML," "Shields," and "PNNL_lq") are more prominent at lower latitudes, where they identify a relatively larger number of ARs than most methods (the machine learning method, "TDA_ML", is an outlier here). The "*Guan_Waliser"* method is another exception owing to its steady rise with latitude throughout the North American coast, which appears more dramatic when normalized precisely because it is so gradual in absolute terms. These normalized results can be further examined by clustering methods according to key differentiating criteria such as whether *absolute* or *relative* thresholds are used (Fig. 8). Focusing on the median of the *absolute* and *relative* clusters (the thick black and blue lines, respectively) reveals excellent agreement in the distribution of AR frequency.

These results suggest that the ARTMIP methods are not identifying fundamentally different features (as could be inferred from the non-normalized results), but rather that their numbers are simply scaled as a function of how restrictive their criteria are. To investigate this further, Fig. 9 presents, for each method, composites of IVT magnitude and identified ARs anytime that method identifies an AR at a point along the northern California coast (39°N, 123.75°W). There are notable differences: less restrictive methods are characterized by a smaller composited AR because they identify both weak and strong events (e.g., "Guan_Waliser", "Rutz", and "Tempest"), whereas more restrictive methods are characterized by a larger composited AR because they identify only strong events (e.g., "CONNECT700" and "PNNL_LQ"). In addition, while most methods' composite ARs exhibit a west/southwest to east/northeast orientation, those based entirely or partly on IWV have a more zonal orientation (e.g., "Goldenson" and "Shields"). However, the methods' composited AR footprints generally cover the same region. The bottom-right panel in Fig. 9 highlights this aspect—anytime the "PNNL_LQ" method, one of the most restrictive, identifies an AR at the coastal point, the number of other methods identifying ARs within the domain shown are counted, and the average over all "PNNL_LQ" ARs is shown in this panel. The results indicate that when one of the most restrictive methods identifies an AR at the coastal point, most other methods (~15), which are less restrictive, also identify an AR near this point, and this number decreases with distance as a function of decreasing IVT. Hence, more restrictive methods' AR composites are shown to be an approximate subset of their less restrictive counterparts.

There are advantages and disadvantages to this normalization approach. One key advantage is that baselining the distribution of AR frequency along these transects is necessary to assess changes predicted by climate models; these results increase confidence in the general shape of the latitudinal distribution. One key disadvantage is that AR-related impacts cannot simply be normalized—emergency management is much more interested in how often these impacts will be encountered than in the general shape of AR frequency along the coast. Thus, more work is needed to constrain the range of AR frequency depicted above, and this is discussed in more detail in Section 5.

3.2 AR Duration

This section discusses AR duration, which is defined as the continuous length of time that a given location is experiencing AR conditions (i.e., is located within the spatial footprint of an AR). For example, if a given method produces an AR duration of 10 h at some location, it means that when this location is within the spatial footprint of ARs, as identified by that method, the average duration of such conditions is 10 h from January 1980 through June 2017, inclusive of all months. AR duration along the North American and European West Coasts varies as a function of method used, but not as greatly as AR frequency (Fig. 10). Focusing on the North American West Coast, most methods exhibit a gradual increase in AR duration from 32°N towards a maximum near 42–44°N (matching the maximum in AR frequency), followed by a steadier decrease northward toward 50°N, plateauing north of there (Fig. 10, top). In their pre-ARTMIP study, focused on Bodega Bay (~38°N along the northern California coast), Ralph et al. (2018b) showed that a group of methods featuring lower IVT thresholds ("Rutz," "Guan_Waliser," and "Gershunov") clusters strongly in both frequency (~23 events per year) and mean event duration (~24 h). Fig. 4 of Ralph et al. (2018b) shows that this agreement, in terms of the number of events (which can be related to AR frequency, given the similar duration), is primarily the product of the fortuitous latitude at Bodega Bay, where results from "Guan_Waliser" cross over with those of "Rutz" and "Gershunov" (Fig. 6). Fig. 10 also shows that these two methods agree on duration at this latitude. More restrictive methods, which typically detect fewer ARs, may also translate into ARs having shorter *average* durations (e.g., "CONNECT700," "TDA_ML,", "Shields and Kiehl"). However, this is not necessarily the case for specific events, as can be seen by examining the AR identification time series shown later in Fig. 12 (top panels).

In general, both less restrictive method criteria and *absolute* thresholds lead to greater AR duration, with the "*Rutz*" and "*Gershunov*" methods described above being among the largest along this transect (the latitude-independent "Lora_global" following closely behind). The "Guan_Waliser" method also produces large AR durations, including the largest south of ~37°N. In addition, the "Guan_Waliser" method also produces the largest mean AR duration (as it does mean AR frequency) through interior western North America (Fig. 10, center). These characteristics arise from the fact that it is a *relative* method that uses percentile-based, latitude- and longitude-dependent IVT thresholds, as explained earlier. In fact, north of ~46°N, this method produces slightly larger mean durations along the interior transect than it does at similar latitudes along the coastal transect, perhaps because its criteria preferentially select for more powerful (and hence, longer-lived) events over regions where IVT is

climatologically weaker. In Europe, the "Lavers" and "Ramos" methods cluster closely together, which is interesting since these two methods were developed within this region (Fig. 10, bottom). Over this region, for many methods, there is little change or a slight decrease in mean event duration as a function of increasing latitude, although the less restrictive "Rutz" method is an exception, peaking at 33 h near 50°N. As with AR frequency, machine learning methods tend to cluster toward lower values of AR duration.

The mean AR duration, after each method is normalized from 0 to 1 and then clustered, shows that the relative distributions of most methods are in good agreement along the North American and European West Coast (Fig. 11). This agreement, however, is not as robust as that observed for AR frequency. For the North American West Coast, *relative* methods produce relatively greater durations between 35–43°N and *absolute* methods produce relatively greater durations between 47–55°—a similar pattern is observed along the European West Coast. These relative differences as a function of latitude are not as apparent for AR frequency (Fig. 7), suggesting that while *relative* and *absolute* methods share similar distributions in overall AR activity, *relative* methods tend to observe longer duration events further south.

3.3. AR Concurrence

This section analyzes the extent to which the ARTMIP methods agree or disagree on the identification of AR conditions along the North American West Coast during events of varying intensity, and the relationship between the methods' identification of ARs conditions and observed precipitation. To do so, the methods' identification of AR conditions along a selected coastal transect during two events, one strong and one weak, are explored. It is important to note that this analysis and the results shown in Fig. 12 are based on the peak IVT (blue time series in top panel) and the presence (or absence) of AR conditions (black dots in top panel) along the entire coastal transect, and not at an individual point.

The first event (12–16 February 2014) is characterized by a broad area of IVT $\geq$ 250 kg m^{-1} s^{-1} making landfall along the U.S. West Coast and extending inland, with areas exceeding the 85th percentile of climatological IVT embedded within the core (Fig. 12a). This event produced heavy precipitation along the coastal and the interior northwestern U.S., triggering a series of avalanches that resulted in 10 fatalities (Hatchett et al. [2017]). Most of the ARTMIP methods identify AR conditions along the coastal transect either throughout, or nearly throughout, the entire period. Some methods, such as "Brands_v2", "Connect500", and "Payne", are very sensitive to periodic surges and lulls in IVT magnitude, identifying ARs during the former. Methods that are more restrictive, such as "Connect700," "PNNL1_hagos," "PNNL_lq," and "TDA_ML" do not identify AR conditions as frequently, particularly at times when peak IVT along the coast drops below their more restrictive thresholds. Furthermore, it must be noted that some methods, such as the "PNNL" methods above, only identify ARs if and when they intersect the coast, but not before or after.

The second event (23–24 October 2006) is characterized by a broad area of IVT $\geq$ 250 kg m^{-1} s^{-1} terminating along the coast of British Columbia, with the 250 kg m^{-1} s^{-1} contour

overlapping the northern edge of the selected transect (Fig. 12b). Precipitation is very light, and the authors are not aware of significant impacts associated with this event. The ARTMIP methods generally disagree as to whether or not AR conditions occur along this transect, and the disagreement extends beyond differentiation into *absolute* and *relative* methods. The "Rutz" and "Tempest" methods most frequently identify an AR along the transect during this time period—both are based on an absolute threshold of IVT $\geq$ 250 kg m^{-1} s^{-1}, but the "Rutz" method utilizes an Eulerian framework for identifying ARs, whereas the "Tempest" method utilizes a Lagrangian framework. Certain *relative* methods such as "Guan_Waliser," "Lora," and "Shields" also identify an AR along this transect nearly 50% of the time (or during adjacent time steps). Other methods, such as "PNNL1_hagos" and "PNNL2_lq", only identify instances when AR conditions are met at the coast (i.e., landfalling ARs) so very few instances were denoted as AR conditions for this AR that barely made landfall along the coast of British Columbia.

One key point is that there is a difference between identifying AR conditions at one time step and identifying an AR event, which is often defined as having some minimum duration (such as the 12-h minimum in Fig. 10). AR duration plays a key role in storm-total precipitation and streamflow (Ralph et al. 2013), and along with peak IVT intensity, forms the basis of a forthcoming AR scale, which categorizes the strength and impacts of ARs (Ralph et al. 2019). A further consideration is that while prolonged AR duration often drives impacts over land, the peaks and troughs in IVT intensity along the coast are of great interest to those focusing on the physical processes involved in strengthening, maintaining, or weakening ARs. Finally, based on this limited analysis, ARTMIP methods exhibit greater agreement regarding those storms that are more meteorologically impressive and associated with heavier precipitation. Of course, significant meteorological events do not always produce significant impacts, and many factors need to be considered, but it is encouraging that all else held constant, a large majority of methods agree to classify "the big ones" as ARs.

3.4 AR Seasonality

In this section, we assess AR seasonality by calculating, at each latitude, the number of methods that yield a maximum AR frequency during a given month (Fig. 13; i.e., for any latitude, the sum across all columns will be equal to the number of methods for that transect). The month of maximum AR frequency along the North American West Coast is characterized by a gradual shift from north to south during the course of the boreal cool season (Fig. 13, top). More specifically, it occurs between 48–54°N (~British Columbia and Vancouver Island) during October, 42–48°N (Washington and Oregon) during November, 36–42°N (northern California) during December, and south of 36°N (southern California) during January. This result agrees well with the results for peak IVT intensity shown by Dettinger et al. (2018), among others. The month of maximum AR frequency should not be confused with AR frequency—in other words, the blue shading indicating a December maximum near 37°N means that a large majority of methods agree that AR frequency, at this latitude, features a maximum in December. It does not mean that the December AR frequency is greater here than some other location, though it might be.

Along the interior western North America transect, the month of maximum AR frequency is quite varied as a function of latitude (Fig. 13, center). North of ~45°N, the pattern is similar to that along the coast, with the month of maximum AR frequency gradually shifting southward September through November. South of ~41°N, this southward shift continues, to some extent, into December and January, but is less clearly seen because of the influence of the monsoon circulation, which produces a maxima during September and October at these latitudes. Perhaps the most unexpected phenomenon shown here is the June maximum between 41–45°N. At these latitudes, during June, seasonally increasing moisture interacts with a jet stream that remains sufficiently strong to produce a relatively large number of ARs. However, given a warmer atmosphere in June than during winter months, it is likely that many of these ARs fail to fully saturate the atmospheric column and produce less precipitation than ARs of similar IVT magnitude during the winter, particularly over lower elevations.

The month of maximum AR frequency along the European West Coast is characterized by a rapid shift from north to south concentrated during September–October, at the onset of the boreal cool season (Fig. 13, bottom). More specifically, it is maximized north of 55°N during September and south of 55°N during October (similar to the seasonality of these same latitudes along the North American West Coast; Gershunov et al. [2017]). At some latitudes, such as those near Scotland or southern Spain, some methods identify December as the month of maximum AR frequency. One possible explanation is that these latitudes can observe landfalling ARs from a greater variety of directions (see the inset maps of grid points used for coastal transects), and the tendency to do so may vary by month. It is also possible that some of this is due to a more fragmented European West Coast versus that of North America. For both the North American and European West Coasts, it is notable that at least 1 or 2 methods identify January as the month of maximum AR frequency at nearly every latitude. This results from a wide variety of methods identifying January along the North American West Coast, and the "Lavers" (further south) and "Shields" (further north) methods consistently identifying January along the European West Coast.

3.5 AR Zonal-Mean Area, Poleward IVT, and "Efficiency"

One oft-quoted result, developed from an early series of seminal papers on ARs (Newell et al. 1992; Zhu and Newell 1998), is that ARs are responsible for ~90% of poleward water vapor transport in the midlatitudes, despite encompassing only ~10% of global circumference at any given latitude and time. This section derives motivation from this early work, and explores related metrics across various ARTMIP methods. Results shown are limited to *global* methods, which can be compared to each other because they consider all latitudes and longitudes. In contrast, *regional* methods cannot be compared because they consider only certain regions.

Zonal-Mean Area

The first metric examined here is the zonal-mean AR area (i.e., the time-mean spatial footprint, along a given latitude band, of identified or tracked ARs), expressed as a fraction of

global circumference, for each *global* method (Fig. 14, top). Most of the *global* methods are characterized by a maximum zonal-mean AR area in the midlatitudes (~10% of global circumference), a rapid decrease toward higher latitudes, and a gradual decrease toward lower latitudes. The rapid decrease toward higher latitudes is due to rapidly decreasing mean water vapor (and hence, IVT) at these latitudes, whereas the gradual decrease towards lower latitudes is dominated by decreasing mean wind (and hence, IVT) further from the mean storm track. One notable exception is the "Rutz" method, which identifies a large fraction of the inter-tropical convergence zone as an AR, since it does not account for climatology and has no width requirement. In general, *absolute* methods exhibit greater variance in zonal-mean AR areas as a function of latitude than *relative* methods. In addition, less restrictive methods predictably identify greater zonal-mean AR areas overall than more restrictive methods.

Here, the "*Guan_waliser*" method is interesting in two respects. First, it identifies larger zonal-mean AR areas at higher latitudes of both hemispheres than any other method. This arises because this method uses a latitude-dependent 85th percentile IVT threshold, where the direct influence on AR activity from the climatological meridional gradient in IVT tends to be smoothed out, resulting in a more gradual decrease in AR occurrence toward high latitudes, as also explained earlier. Second, for the same reason above, the fractional zonal-mean AR area identified is remarkably stable throughout the mid latitudes of both hemispheres at nearly 10%, which is the value from Zhu and Newell (1998) for the fraction of global circumference encompassed by ARs. A number of other methods (i.e., "*Rutz*," "*Lora_global*," "*Lora_npac*," "Tempest," "*Mundhenk*," and *"Connect500*") also approach this value at low or mid latitudes, but their distributions are more variable as a function of latitude.

The midlatitude (~30–60°N/S) global circumference occupied by ARs is in the range of ~2–15% when considering all *global* methods, excluding the "CONNECT" methods (for which the range is lower, due to their fairly restrictive thresholds). This is important for three reasons. First, being significantly below 50% means that even the least restrictive methods examined here (such as "Rutz") are identifying discrete features that are making large contributions to IVT relative to their size. Second, the average global circumference occupied in the core of the midlatitudes (~45°N/S) being ~5–15% means that these features occupy more space than the cold-frontal zones associated with extratropical cyclones, and therefore the concept of an AR is distinct and useful. Finally, it is encouraging that these results, based on 5 methods and ~38 years of global data, align so well with those discussed by Zhu and Newell 20 years ago.

Zonal-Mean Poleward IVT

The second metric examined here is the AR-related zonal-mean poleward IVT (i.e., poleward IVT occurring within the spatial area of ARs; Fig. 14, middle). For all methods, the AR-related zonal-mean poleward IVT is maximized in the midlatitudes. The variation in the magnitude of its maximum and meridional range clearly exhibits a dependence on the threshold magnitudes chosen for identifying ARs, which is most clearly seen using the

absolute methods. For example, "*connect500*" attributes a greater amount of zonal-mean poleward IVT to ARs than "*connect700*," simply because the former uses an IVT threshold of 500 (versus 700) kg m^{-1} s^{-1}, and identifies ARs as spatially larger features. Similarly, the "tempest" and "rutz" methods, based on 250 kg m^{-1} s^{-1}, attribute an even greater amount of zonal-mean poleward IVT to ARs, with "rutz" attributing more than "tempest" because it is less restrictive with other criteria. The *relative* "lora" method, with a less restrictive threshold requirement of IVT ≥ 100 kg m^{-1} s^{-1} above climatology, attributes nearly as much zonal-mean poleward IVT to ARs as "rutz." The other relative methods "guan_waliser" and "mundhenk" have more restrictive criteria, and attribute smaller fractions of zonal-mean poleward IVT to ARs. The "*guan_waliser*" method is notable because of its relative smoothness at high latitudes, being the most generous in attributing zonal-mean poleward IVT to ARs north of ~55°N. It is interesting that this is roughly the same latitude at which the AR frequency of the "guan_waliser" method becomes greater than that of "rutz" and "lora" along the North American and European West Coasts (Fig. 6).

Zonal-Mean AR Efficiency

The final metric examined in this section is the zonal-mean AR "efficiency," defined, in terms of the two metrics examined previously, as the ratio of zonal-mean poleward IVT to the fractional (i.e., unitless) zonal-mean spatial area of ARs. It is referred to here as the zonal-mean AR efficiency because it describes the quantity of poleward water vapor transport per unit area of AR. One inherent problem with this metric is that efficiency, as defined above, is naturally higher for more restrictive AR identification and tracking methods. However, it is still interesting to explore this metric, and particularly how it changes as a function of latitude.

The methods with the largest zonal-mean AR efficiency across most latitudes are "CONNECT500" and "CONNECT700" (Fig. 14, bottom). This is not surprising given that the fairly restrictive criteria of IVT ≥ 500 and 700 kg m^{-1} s^{-1} limits the number of ARs identified by these methods to only the strongest of those identified by other methods. In fact, the lack of events due to these restrictive criteria is clearly seen to affect the results over high latitudes. The other methods cluster more closely together, particularly over mid and high latitudes, although the "guan_waliser" method identifies more ARs over Antarctica, owing to its less restrictive criteria, and hence the efficiency is lower. In the tropics, the "rutz" method, which was designed for mid-latitude applications, is the least efficient, as it often identifies regions of broad tropical moisture transport as ARs. In contrast, the "lora" method becomes more efficient at these latitudes, since it requires that IVT exceed the climatological mean by 100 kg m^{-1} s^{-1}. Finally, AR efficiency using the "mundhenk" and "tempest" methods, which are based on IVT ≥ 94^{th} percentile of the anomalies above the climatology and IVT ≥ 250 kg m^{-1} s^{-1}, respectively, are relatively steady across all latitudes.

In summary, this section shows that most global methods used within ARTMIP broadly reproduce the classic results of Zhu and Newell (1998) in terms of AR size and significance for global water vapor transport.

3.6 Spread among Methods

This section explores the relative difference between results for each individual method and the all-method median for AR frequency, the month of maximum AR frequency, and the seasonal range of AR frequency (Fig. 15). Results are presented along 7 selected transects: the Pacific Northwest (PNW; 41–52.5°N), northern California (NorCal; 35–41°N), southern California (SoCal; 32–35°N), the interior western U.S. (WUS_In; 32–54°N), South America (SAmer; 18–56°S), the United Kingdom (UK; 49–60°N), and Iberia (Ib; 35–48°N). This analysis is fairly exhaustive, and a full description of every aspect would be very lengthy, so this section focuses on the highlights. This analysis offers some insight as to which methods produce results closer to the median, and which methods produce results further from the median, perhaps even being characterized as outliers. Here, relative difference for a given region and given method are calculated as the difference between a given method and the all-method median normalized by the all-method median.

For AR frequency, the methods closest (lighter color shading) to the all-method median (<±36%) are "Goldenson," "Mundhenk," "Payne," and "Shields," whereas the methods furthest (darker color shading) from the all-method median (>±60%) are "CONNECT700," "Gershunov," "Guan_Waliser," "PNNL2_lq," and "Rutz" (Fig. 15, top). Generally, the methods closest to the all-method median are *relative* methods, with the exception of "Goldenson," whereas the methods furthest from the all-method median are *absolute* methods, with the exception of "Guan_Waliser." This breakdown by absolute/relative is not too surprising since absolute methods tend to accentuate climatological differences while relative methods tend to diminish them. Of those methods furthest from the median, the two methods identifying much lower AR frequency are "CONNECT700," which uses a high IVT threshold of 700 kg m^{-1} s^{-1}, and "PNNL2_lq," which has a number of restrictive criteria (Table 1). The three methods identifying much higher AR frequency are "Gershunov" and "Rutz," which have similar and less restrictive criteria, and "Guan_Waliser," which also has fairly low criteria (85% climatological IVT) for a *relative* method. It is worth noting that most methods are relatively far from the median for AR frequency along the transect through the interior western U.S., which results partly from differences in how methods assess ARs over complex terrain, and partly from calculating percentage differences between small numbers. Another point of interest is the rather notable differences in the three "Brands" methods, which shows that relatively minor changes to the AR identification method (e.g., the threshold percentiles) can significantly alter results. Finally, the regional methods "Lavers," "Ramos," and "Viale," which are available over the UK, Iberia, and Chile, respectively, all produce a lower AR frequency than the median (i.e., global methods) over these regions. One possible explanation for this interesting result is that regional methods are more finely tuned to their respective areas, and that this is manifested as more restrictive criteria. For example, both "Lavers" and "Ramos" use time-dependent percentiles based on climatological IVT, and "Viale" imposes the restriction that an AR must be associated with a frontal system. The machine learning technique, 'TDA_ML', produces AR frequencies below the median, possibly because the algorithm employed by Muszynski et al. (2019) exhibits a fairly strong resolution-dependent decrease in the 'sensitivity score' (the proportion of

identified ARs that are correctly identified) as resolution increases, with over 25% of features being mis-classified (relative to their training dataset) as non-ARs at high resolution. Muszynski et al. (2019) hypothesize that this is due to an interaction between the decrease in smoothness of the IWV field as resolution increases and the underlying topology-based method that they use to identify potential ARs.

The month of maximum AR frequency is also examined (Fig. 15, middle), and the following examples assist with interpretation: the median month of maximum AR frequency over the UK is September (i.e., peak month of 9), whereas for "CONNECT700," it is October (i.e., peak month difference of 1). Similarly, the median month of maximum AR frequency over SoCal is February (i.e., peak month of 2), whereas for "Goldenson," it is December (i.e., peak month difference of -2). In many cases, there is no difference between the all-method median and the month of maximum AR frequency identified by most methods, though there are some notable exceptions. For SAmer, both the "Guan_Waliser" and "Viale" methods, the latter of which is focused on this region, feature a month of maximum frequency 4+ months later than the median of February. This effectively means that whereas most methods are identifying the maximum in Austral Summer, these methods identify the maximum in Austral Winter, which is possible because the storm track in this region is less seasonally variable than over the Northern Hemisphere (e.g., Trenberth [1991]). The "Shields" method has a tendency to identify frequency maxima a few months later than the median over nearly every region with the exception of SoCal, and this may be due to the tendency of this method to detect only the stronger storms, typically in mid-winter, when the eddy-driven jet is further south. Only the "Lora" and "Tempest" methods show no deviations from the median over all regions, though a few other methods come close.

The third metric assessed here is the seasonality (i.e., range) of AR frequency (Fig. 15, top), and many results are similar to those for AR frequency itself. The methods closest (lighter color shading) to the all-method median (<±30%) are "Brands_v1," "Goldenson," "Mundhenk," and "Payne," the latter three of which are all close to the median for AR frequency as well (Fig. 15, top). The methods furthest (darker color shading) from the all-method median (>±40%) are "CONNECT700," "PNNL1_hagos," "PNNL2_lq," "Rutz," "Shields," and "TDA_ML." As with AR frequency, the methods closest to the all-method median are *relative* methods, with the exception of "Goldenson," whereas the methods furthest from the all-method median are more mixed. Of those methods furthest from the median, the four methods identifying much weaker seasonality are "CONNECT700" and the "PNNL" methods, which are both quite restrictive, and the machine learning technique, "TDA_ML." The two methods identifying much stronger seasonality are "Rutz" and "Shields," the latter of which features a larger range in AR frequency across most regions, despite having a smaller AR frequency in a few of them. As with AR frequency, most methods are relatively far from the median for the range in AR frequency along the transect through the interior western U.S. Of the global methods, "Lora_global" and "Rutz" exhibit a stronger seasonality over all regions, whereas "CONNECT700" exhibits a weaker seasonality over all regions except SAmer. As with AR frequency, the regional methods "Lavers,"

"Ramos," and "Viale" all produce a weaker seasonality than the median (i.e., global methods) over their respective regions.

For both the AR frequency and the seasonality (i.e., range) of AR frequency, it is notable that each method generally exhibits either a positive or a negative relative difference from the median across all transects. Exceptions to this generalization are most commonly noted along the interior western U.S. transect, which is the only one located amidst complex topography far from a coast. Hence, the ARTMIP methods used to identify ARs do not seem particularly sensitive to the region in which they are employed.

4. Discussion

Results based on the ARTMIP methods have been described in terms of "clusters," which are groupings of methods that approach AR identification and tracking similarly in a few critical ways. These clusters differentiate between methods with very different approaches, and often very different results, without knowing the nuances of each individual method within the cluster. The three cluster pairs this text has focused on, as discussed in Section 2, are *absolute/relative*, *global/regional*, and the more subjectively defined less/more restrictive. The subdivision of *relative* methods into *latitude-dependent* and *latitude-independent relative* methods has also been noted. There are many other cluster pairs (e.g., *length/no length*, or *time-slicing/time-stitching*) that have been omitted from this study in the interest of brevity. In each of these cluster pairs, both clusters feature advantages and disadvantages, some of which are discussed below.

The *absolute* and *relative* clusters are perhaps the most fundamentally different in their approach to identifying ARs. One key advantage of *absolute* methods is that they ensure a minimum physical threshold is met before features are identified as ARs, which can be useful when considering only stronger events. Some methods, such as "CONNECT700" (threshold IVT $\geq$ 700 kg m^{-1} s^{-1}), are designed to consider only the strongest events. Furthermore, ARs identified following an *absolute* method are sometimes comparable across regions, provided the appropriate threshold is carefully chosen to pursue specific applications. In that regard, a relative method might be particularly useful when a single absolute threshold does not work well across all the regions of interest—an example being over polar regions, where a temperature-adjusted (i.e., climatology-dependent) AR threshold has proven useful in detecting AR landfalls (Gorodetskaya et al. 2014). Another good example is the lengthy southwestern coast of South America, which stretches from 18–56°S and encompasses a wide range of climatological IVT values. Some methods combine relative and absolute thresholds to leverage the advantages of each. For example, the *latitude-dependent relative* "Guan_Waliser" method combines a relative threshold (IVT $\geq$ 85th percentile of climatological IVT) with an absolute threshold (IVT $\geq$ 100 kg m^{-1} s^{-1}), the latter of which eliminates extremely weak features, particularly closer to the poles.

Another key advantage of *relative* methods is that they facilitate the pursuit of AR science in regions where it is more difficult to do so using *absolute* methods. For example, imagine one wants to investigate the impacts associated with the inland-penetrating AR depicted in Fig. 2.

The "Guan_Waliser" method may be a good choice since it identifies a broad inland region as being located within the AR (other *relative* methods have more restrictive criteria, and identify less area within the AR), and many impacts within this region could be AR related. Of course, it may be that in some cases this region is too broad, and choosing an *absolute* method that still highlights the inland penetration of the AR, but focuses more closely along its axis or region of core intensity, is appropriate. These are very difficult decisions that need to be made based on the specifics of the question being asked.

Climate change poses yet another point for consideration. As atmospheric temperature and moisture increase following the Clausius-Clapeyron equation, IWV will increase, and IVT will increase (unless increases in water vapor are offset by decreases in wind). As the background moisture field increases, *absolute* methods using thresholds based on the current climatology may struggle to distinguish between this increased background moisture and coherent ARs resulting from dynamical processes. Hence, *relative* methods using thresholds based on climatology may be better suited to assess relative changes in ARs due to dynamic and thermodynamic factors between our current climate and that of the future. For example, one can compare two relative methods – one in which the percentile threshold is applied to the respective climatology of the present and future (to isolate the dynamic factor) and one in which the percentile threshold is applied to the present climatology and then the corresponding absolute threshold is used in the future (hence including both thermodynamic and dynamic factors) – to separate the dynamic and thermodynamic effects. On the other hand, impacts are generally not considered in relative terms, and one must be careful in this regard. Forthcoming work by the ARTMIP community will address this issue in depth by examining ARTMIP methods under future climate scenarios, and data processing is already underway.

Another important set of clusters examined in this study is that of *global* and *regional* methods. One key advantage of *global* methods is simply the global coverage of results, unlike *regional* methods, which are limited. Another, more speculative, advantage of *global* methods is that their development may benefit from using a global perspective rather than a focus on one region, where ARs may frequently take on characteristics not observed in most locales. In contrast, one key advantage of *regional* methods is that they are specifically tuned to ARs and AR-related impacts over a specific region and hence may be the most useful for answering key science questions particular to those regions.

A final distinction made, qualitatively, throughout this study is that between less restrictive and more restrictive methods—a very subjective distinction based on their criteria, and usually only useful when comparing one method to another. One key advantage of less restrictive methods is that they facilitate AR science and impacts in regions where ARs are very rare using more restrictive methods (e.g., the usefulness of "Gorodetskaya" over polar regions, and "Guan_Waliser" or "Rutz" over continental interiors). Of course, the disadvantage is the reverse—less restrictive methods may result in the attribution of impacts to ARs, when in fact the associated dynamics and vapor fluxes are very weak, or merely remnants of a once-robust AR. Researchers need to carefully weigh the advantages and disadvantages of their approach to answering a given scientific question. It is important to

remember that in this paper, distinctions between less restrictive and more restrictive methods can really only be made within a cluster (work is underway to more objectively quantify restrictiveness across clusters). For example, Fig. 6 (middle) makes every method seem restrictive in comparison to the less restrictive *relative* "Guan_Waliser" method, when in fact there are also more restrictive *relative* methods (e.g., "Payne"), less restrictive *absolute* methods (e.g., "Rutz") and more restrictive *absolute* methods (e.g., "Tempest"). In the case of "Tempest", greater restrictiveness arises because objects are required to remain long and thin across time, whereas ARs tend to spread out as they encounter land. Finally, one key advantage of using more restrictive methods is that they highlight only the strongest events, which will likely (though not always) produce the most severe impacts.

A number of ARTMIP methods are based on the identification of features meeting certain geometric criteria, throughout which either IVT $\geq$ 250 kg m^{-1} s^{-1} (e.g., "Brands," "Gershunov," "Rutz," and "Tempest") or IWV $\geq$ 20 mm (e.g., "Goldenson," "Hagos," "Ralph," and "Wick"). In addition, the recently developed AR scale, described by Ralph et al. (2019), establishes IVT $\geq$ 250 kg m^{-1} s^{-1} as the minimum threshold required to categorize an event as an AR. The AR science community increasingly recognizes the importance of water vapor transport within ARs and now strongly favors IVT over IWV for diagnosing such features. Therefore, IVT $\geq$ 250 kg m^{-1} s^{-1} seems to be a reasonable starting point. However, there are many cases in which IVT of this magnitude will be primarily beneficial, and it becomes worthwhile to identify only stronger, more hazardous ARs. This is one rationale for higher minimum thresholds such as 500 or 700 kg m^{-1} s^{-1}, as used in "CONNECT500" and "CONNECT700," respectively. The rationale for a higher minimum threshold can also be climatologically and/or regionally based, as is the case for the very high background moisture field over the southeastern U.S., a region in which 500 kg m^{-1} s^{-1} was used by Mahoney et al. (2016; results not available for this study). It should also be noted that IVT thresholds below 250 kg m^{-1} s^{-1} can be useful both in regions with climatologically lower IVT, and in cases where long-duration, low-intensity IVT events may produce significant impacts.

A number of *relative* methods use thresholds based on IVT $\geq$ 85th percentile of climatological IVT, along with an absolute IVT threshold that serves as a floor, or minimum IVT requirement, to identify features as ARs (e.g., "Guan_Waliser," "Lavers," "Payne," "Ramos," and "Viale"). Still other *relative* methods threshold based on IVT exceeding daily climatology by some raw value such as 100 or 250 kg m^{-1} s^{-1} (e.g., "Lora" and "Walton"), or use some other method (e.g. "Gorodetskaya," "Mundhenk," and "Shields"). Among these, "Guan_Waliser" is the least restrictive due to a minimum IVT requirement of only 100 kg m^{-1} s^{-1}, and this causes its AR frequency (and results directly associated with AR frequency) to be clear outliers in polar regions and continental interiors where IVT is climatologically low. That said, it is an extremely useful outlier, because it often identifies regions downstream of mountain barriers as within an AR, whereas most other methods do not. It can be argued, based on the AMS Glossary definition, that these regions are not necessarily located within the spatial footprint of an AR. However, the usefulness is found in identifying and attributing impacts to the ARs likely responsible for them, even if the spatial footprints of these ARs, and their impacts, do not directly overlap. The "Gorodetskaya" method, from which results

were not available for this study, also identifies ARs (and AR-related impacts) in regions of very low IWV/IVT, having been designed specifically to identify intrusions of anomalously moist air into polar regions. To be more consistent with the AMS Glossary definition, such features could potentially be described as "decaying" ARs, or by some other term, which indicates that they are no longer associated with the extratropical cyclones and/or dynamic processes critical to their genesis.

Quantifying the uncertainty in AR-related impacts (and how they may change in the future), most of which are in some way related to precipitation, is a major motivation behind ARTMIP. Some sense of impacts can likely be inferred from the results for AR climatology highlighted in this study. However, a more complete assessment of the advantages and disadvantages associated with individual methods and with certain clusters will be possible only after some quantification of the uncertainty in AR-related precipitation takes place. Future ARTMIP work plans to address this subject.

Finally, ARTMIP expects to produce a number of new results and publications over the coming years. The most salient of these is a pair of Tier 2 summary papers, which will present results from all ARTMIP methods applied to output from a high resolution version of the Community Atmospheric Model (Wehner et al. 2014) and available CMIP5 models under historical and RCP8.5 forcing scenarios. Numerous studies (e.g., Warner et al. [2015]; Gao et al. [2016]; Shields and Kiehl [2016a,b]; Espinoza et al. [2018]) have already examined changes in AR climate and impacts under climate change scenarios, but, as with studies of current AR climate and impacts, these suffer from uncertainty that arises due to the usage of different AR identification and tracking methods. The Tier 2 summary papers will quantify these uncertainties. In addition, ARTMIP participants have already planned a number of studies on topics ranging from quantifying differences in ARs based on the reanalysis product used to trends in ARs over time, and a variety of other topics.

5. Recommendations

The results presented in this study indicate a large degree of uncertainty in the climatological characteristics of ARs resulting from differences in the methods used to identify and track them. This uncertainty is reduced within "clusters" of methods that share similar approaches to AR identification and tracking, but even then, uncertainty arises due to differences in thresholding variable and magnitude, geometric considerations, and other criteria. As stated in the introduction, this should not be surprising—each method was developed to answer a different question, and different answers naturally arise. This diversity benefits the community in that it offers a wide variety of approaches to answering new questions that may arise. Nevertheless, the AR science community will be interested in recommendations regarding which of these methods or clusters best answers their questions.

Here, the authors provide generalized recommendations regarding the types of AR identification and tracking methods that are most advantageous for certain applications, and ideas regarding future method development.

The authors generally recommend *absolute* methods for studies focused on the relationship between ARs and large-scale atmospheric patterns, dynamic processes, and physical mechanisms for our current climate in the mid-latitudes (~30–60°N/S). The AR science community increasingly recognizes these features, particularly as they relate to extratropical cyclones, as playing a key role in defining the AR, as well as intense water vapor transport (see the AMS Glossary definition). These methods have the advantage of being based on fixed, observable thresholds (e.g., 20 mm of IWV or 250 kg m^{-1} s^{-1} of IVT), which are well suited to answer questions related to AR dynamics such as their growth, maintenance, and decay. Hence, these methods are preferred at midlatitudes, where the dynamics associated with extratropical cyclones typically drive the life cycle of ARs. *Absolute* methods are also recommended for most weather forecasting applications because they are straightforward and intuitive, though forecasts based on anomalies, percentiles, or return intervals can be very effective in communicating to more knowledgeable audiences. Finally, it must be remembered that AR-related impacts can occur well outside the spatial footprint outlined by *absolute* methods, particularly when the threshold used is more restrictive.

The authors generally recommend *relative* methods for studies focused on attributing a wide variety of hazards (e.g., heavy rainfall, flooding, and wind) to ARs. These methods have the advantage, as was described at length above in reference to the "Guan_Waliser" method, of placing regions far from the core of an AR within the spatial footprint of the AR, which facilitates the attribution of impacts within that footprint to the AR itself. These methods are preferred at tropical and subtropical latitudes, because they often more effectively filter the broad regions of IVT that occur throughout the tropics. They are also preferred at polar latitudes, because they often identify features that are climatologically anomalous, despite having low IWV or IVT values relative to mid-latitude features. Hence, these methods are well suited to answer questions related to the occurrence and impacts of climatologically anomalous moisture surges around the globe, but it must be remembered that some results may not translate well from one region to another.

The authors, at this time, defer recommendations regarding studies involving climate change until the Tier 2 analyses have been completed. It is likely that there will be advantages and disadvantages to both absolute and relative methods, just as there will be to both sides of any number of other cluster pairs.

The authors recommend more restrictive methods for studies focusing on dynamic processes related to the core of the AR and for studies focused on a subset of generally stronger ARs, which more restrictive methods will select. For studies related to the attribution of precipitation and other impacts to ARs, the authors recommend carefully considering the goals and objectives of each study. One advantage of less restrictive methods is that identified ARs are associated with larger spatial footprints, which aid in evaluating all impacts potentially related to ARs. However, these larger footprints can cause the average impacts associated with ARs to be quite low, which could prove misleading to the public. Hence, more restrictive methods may be more suitable for highlighting the extreme impacts that occur typically along the axis of greatest IVT within an AR.

The authors also recommend that studies focusing on a particular region should consider basing their analyses on regional methods developed to assess that region. These methods may take into account important regional characteristics of ARs and their impacts, whereas those focused on other regions, or global methods, may not. Of course, previously developed methods will not be sufficient to answer certain questions, and yet other studies might benefit from using global methods, which facilitate comparisons between different parts of the world.

The authors recommend a few directions for future work in the area of AR identification and tracking method development that would benefit the AR science community. The AMS Glossary defines an AR as "a long, narrow, and transient corridor of strong horizontal water vapor transport that is typically associated with a low-level jet stream ahead of the cold front of an extratropical cyclone". The AR science community is increasingly recognizing the key role of dynamic processes in terms of defining the AR (personal correspondence with participants and observation of presentations at 2018 International Atmospheric Rivers Conference). Hence, the first recommendation is to emphasize methods based on IVT, which incorporates wind as well as moisture, over those based on IWV. In this regard, wind serves as a proxy variable for a number of dynamic processes, and is fundamental to the moisture transport associated with ARs. The second recommendation is development of an interactive online tool that allows researchers to compare multiple methods, along with other relevant layers (e.g., IWV, IVT, geopotential height, temperature, and precipitation), in real time. This tool would assist researchers in determining which methods are most useful for their specific applications. The third recommendation is that future work should consult the ARTMIP archive and literature provided online. This provides future studies with the contextual background of what methods already exist, which studies have been performed, and how new results best fit into this emerging field of study.

ARTMIP has produced, and will continue to produce, an astonishing quantity of data that can be mined to improve our understanding of ARs and their impacts. The authors anticipate, and indeed already plan, a number of studies that will address various topics, and others are encouraged to do the same. The generalized recommendations above are the authors' best guidance at this time, but they are by no means a panacea. While we think that these recommendations are useful, the history of science suggests that the most interesting results will arise only when they are ignored. We welcome those developments.

Acknowledgments

ARTMIP is a grass-roots community effort and includes a collection of international researchers from universities, laboratories, and agencies. Co-chairs and committee members include Jonathan Rutz, Christine Shields, L. Ruby Leung, F. Martin Ralph, Michael Wehner, Ashley Payne, and Travis O'Brien. Details on catalogues developers can be found on the ARTMIP website. ARTMIP has received support from the US Department of Energy Office of Science Biological and Environmental Research (BER) as part of the Regional and Global Climate Modeling program, and the Center for Western Weather and Water Extremes (CW3E) at Scripps Institute for Oceanography at the University of California, San Diego.

Data Availability

All ARTMIP data, (including the MERRA-2 source data), is available from the Climate Data Gateway, DOI:10.5065/D6R78D1M (ARTMIP Tier 1 Catalogues), DOI: 10.5065/D62R3QFS (MERRA-2 source data).

References

Albano, C. M., M. D. Dettinger, and C. E. Soulard (2017), Influence of atmospheric rivers on vegetation productivity and fire patterns in the southwestern U.S., *J. Geophys. Res. Biogeosci.*, 122, 308–323, doi:10.1002/2016JG003608.

Ault, A. P., C. R. Williams, A. B. White, P. J. Neiman, J. M. Creamean, C. J. Gaston, F. M. Ralph, and K. A. Prather (2011), Detection of Asian dust in California orographic precipitation, *J. Geophys. Res.*, 116, D16205, doi: 10.1029/2010JD015351.

Alexander, M.A., J.D. Scott, D. Swales, M. Hughes, K. Mahoney, and C.A. Smith, 2015: Moisture Pathways into the U.S. Intermountain West Associated with Heavy Winter Precipitation Events. *J. Hydrometeor.,* **16**, 1184–1206, https://doi.org/10.1175/JHM-D-14-0139.1

Brands, S., J.M. Gutiérrez, and D. San-Martín, 2017: Twentieth-century atmospheric river activity along the west coasts of Europe and North America: algorithm formulation, reanalysis uncertainty and links to atmospheric circulation patterns. Clim. Dyn. 48 (9-10), 2771-2795, doi:10.1007/s00382-016-3095-6.

Chen, X., L.R. Leung, Y. Gao, Y. Liu, M. Wigmosta, and M. Richmond (2018), Predictability of extreme precipitation in western U.S. watersheds based on atmospheric river occurrence, intensity, and duration. *Geophys. Res. Lett.*, revised.

Cheng BS, Chang AL, Deck A, Ferner MC. 2016 Atmospheric rivers and the mass mortality of wild oysters: insight into an extreme future? Proc. R. Soc. B 283: 20161462. http://dx.doi.org/10.1098/rspb.2016.1462

Cordeira, J.M., F.M. Ralph, A. Martin, N. Gaggini, J.R. Spackman, P.J. Neiman, J.J. Rutz, and R. Pierce, 2017: Forecasting Atmospheric Rivers during CalWater 2015. *Bull. Amer. Meteor. Soc.,* **98**, 449–459, https://doi.org/10.1175/BAMS-D-15-00245.1

Dettinger, M. D., F. M. Ralph, T. Das, P. J. Neiman, and D. R. Cayan, 2011: Atmospheric rivers, floods, and the water resources of California. *Water,***2011** (3), 445–478.

Dettinger, M.D., F.M. Ralph, and J.J. Rutz, 2018: Empirical Return Periods of the Most Intense Vapor Transports during Historical Atmospheric River Landfalls on the U.S. West Coast. *J. Hydrometeor.,* **19**, 1363–1377, https://doi.org/10.1175/JHM-D-17-0247.1

Espinoza, V., D. E. Waliser, B. Guan, D. A. Lavers, and F. M. Ralph (2018), Global analysis of climate change projection effects on atmospheric rivers, Geophys. Res. Lett., 45, 4299–4308, doi:10.1029/2017GL076968.

Gao, Y., J. Lu, and L.R. Leung, 2016: Uncertainties in Projecting Future Changes in Atmospheric Rivers and Their Impacts on Heavy Precipitation over Europe. *J. Climate,***29**, 6711–6726, https://doi.org/10.1175/JCLI-D-16-0088.1

Gao, Y., J. Lu, L.R. Leung, Q. Yang, S. Hagos, and Y. Qian (2015), Dynamical and thermodynamical modulations on future changes of landfalling atmospheric rivers over western North America. *Geophys. Res. Lett*., 42, 7179–7186, doi:10.1002/2015GL065435.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman,W., Rienecker, M., Schubert, S. D.,Sienkiewicz, M., and Zhao, B.: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), J. Climate, 30, 5419–5454, https://doi.org/10.1175/jcli-d-16-0758.1, 2017.

Gershunov, A, T. Shulgina, F.M. Ralph, D.A. Laver, and J.J Rutz, 2017: Assessing the climate-scale variability of atmospheric rivers affecting western North America, Geophys. Res. Lett., 44, 7900–7908.

Goldenson, N., L.R. Leung, C.M. Bitz, and E. Blanchard-Wrigglesworth, 2018: Influence of Atmospheric Rivers on Mountain Snowpack in the Western United States. *J. Climate*, **31**, 9921–9940, https://doi.org/10.1175/JCLI-D-18-0268.1

Gorodetskaya, I. V., M. Tsukernik, K. Claes, M. F. Ralph, W. D. Neff, and N. P. M. Van Lipzig (2014), The role of atmospheric rivers in anomalous snow accumulation in East Antarctica, *Geophys. Res. Lett.*, 41, 6199–6206, doi: 10.1002/2014GL060881.

Guan, B., and D. E. Waliser (2015), Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies, *J. Geophys. Res. Atmos.*, 120, 12514–12535, doi: 10.1002/2015JD024257.

Hagos, S.M., L.R. Leung, J.-H. Yoon, J. Lu, and Y. Gao (2016), A projection of changes in landfalling atmospheric river frequency and extreme precipitation over western North America from the large ensemble CESM simulations. *Geophys. Res. Lett*., 43, 1357–1363, doi: 10.1002/2015GL067392.

Hatchett, B.J., S. Burak, J.J. Rutz, N.S. Oakley, E.H. Bair, and M.L. Kaplan, 2017: Avalanche Fatalities during Atmospheric River Events in the Western United States. *J. Hydrometeor.*, **18**, 1359–1374, https://doi.org/10.1175/JHM-D-16-0219.1

Huning, L. S., S. A. Margulis, B. Guan, D. E. Waliser, and P. J. Neiman (2017), Implications of detection methods on characterizing atmospheric river contribution to seasonal snowfall across Sierra Nevada, USA, Geophys. Res. Lett., 44, 10445–10453, doi:10.1002/2017GL075201.

Lamjiri, M. A., M.D. Dettinger, F.M. Ralph, and B. Guan (2017), Hourly storm characteristics along the U.S. West Coast: Role of atmospheric rivers in extreme precipitation. *Geophysical Research Letters*, *44*(13), 7020–7028. https://doi.org/10.1002/2017GL074193

Lavers, D. A., G. Villarini, R. P. Allan, E. F. Wood, and A. J. Wade (2012), The detection of atmospheric rivers in atmospheric reanalyses and their links to British winter floods and the large-scale climatic circulation, J. Geophys. Res., 117, D20106, doi:10.1029/2012JD018027.

Lavers, D.A. and G. Villarini, 2013: Atmospheric Rivers and flooding over the central United States. *J. Climate*, **26**, 7829–7836, https://doi.org/10.1175/JCLI-D-13-00212.1

Lavers, D.A., E. Zsoter, D.S. Richardson, and F. Pappenberger, 2017: An Assessment of the ECMWF Extreme Forecast Index for Water Vapor Transport during Boreal Winter. *Wea. Forecasting*, **32**, 1667–1674.

Leung, L.R., and Y. Qian (2009), Atmospheric rivers induced heavy precipitation and flooding in the western U.S. simulated by the WRF regional climate model. *Geophys. Res. Lett.*, 36, L03820, doi:10.1029/2008GL036445.

Lora, J. M., J. L. Mitchell, C. Risi, and A. E. Tripati (2017), North Pacific atmospheric rivers and their influence on western North America at the Last Glacial Maximum, *Geophys. Res. Lett.*, 44, 1051–1059, doi: 10.1002/2016GL071541.

Lukens, K.E., E.H. Berbery, and K.I. Hodges, 2018: The Imprint of Strong-Storm Tracks on Winter Weather in North America. *J. Climate*, **31**, 2057–2074, https://doi.org/10.1175/JCLI-D-17-0420.1

Mahoney, K., D.L. Jackson, P. Neiman, M. Hughes, L. Darby, G. Wick, A. White, E. Sukovich, and R. Cifelli, 2016: Understanding the Role of Atmospheric Rivers in Heavy Precipitation in the Southeast United States. *Mon. Wea. Rev.*, **144**, 1617–1632, https://doi.org/10.1175/MWR-D-15-0279.1

Mudigonda, Mayur, Sookyung Kim, Ankur Mahesh, Samira Kahou, Karthik Kashinath, Dean Williams, Vincent Michalski, Travis A. O'Brien, and Mr Prabhat: Segmenting and Tracking Extreme Climate Events Using Neural Networks. In 31st Conference on Neural Information Processing System, 1–5. Long Beach, CA, USA, 2017.

Mundhenk, B.D., E.A. Barnes, and E.D. Maloney, 2016: All-Season Climatology and Variability of Atmospheric River Frequencies over the North Pacific. *J. Climate*, **29**,4885–4903, https://doi.org/10.1175/JCLI-D-15-0655.1

Muszynski, G., Kashinath, K., Kurlin, V., Wehner, M., and Prabhat: Topological data analysis and machine learning for recognizing atmospheric river patterns in large climate datasets, Geosci. Model Dev., 12, 613–628, https://doi.org/10.5194/gmd-12-613-2019, 2019.

Neiman, P.J., F.M. Ralph, G.A. Wick, J.D. Lundquist, and M.D. Dettinger, 2008: Meteorological Characteristics and Overland Precipitation Impacts of Atmospheric Rivers Affecting the West Coast of North America Based on Eight Years of SSM/I Satellite Observations. *J. Hydrometeor.*, **9**, 22–47, https://doi.org/10.1175/2007JHM855.1

Newell, R.E., N.E. Newell, Y. Zhu, and C. Scott, 1992: Tropospheric rivers? – A pilot study, Geophys. Res. Lett., 24, 2401–2404.

Oakley, N.S., Lancaster, J.T., Kaplan, M.L. et al. Nat Hazards (2017) 88: 327. https://doi.org/10.1007/s11069-017-2867-6

Payne, A.E. and G. Magnusdottir, 2014: Dynamics of Landfalling Atmospheric Rivers over the North Pacific in 30 Years of MERRA Reanalysis. *J. Climate*, **27**, 7133–7150,https://doi.org/10.1175/JCLI-D-14-00034.1

Payne, A. E., and G. Magnusdottir (2015), An evaluation of atmospheric rivers over the North Pacific in CMIP5 and their response to warming under RCP 8.5, *J. Geophys. Res. Atmos.*, 120, 11,173–11,190, doi:10.1002/2015JD023586.

Radić, V., Cannon, A. J., Menounos, B., and Gi, N. (2015), Future changes in autumn atmospheric river events in British Columbia, Canada, as projected by CMIP5 global climate models, *J. Geophys. Res. Atmos.*, 120, 9279– 9302, doi:10.1002/2015JD023279.

Ramos, A.M., R.M. Trigo, M.L. Liberato, and R. Tomé, 2015: Daily Precipitation Extreme Events in the Iberian Peninsula and Its Association with Atmospheric Rivers. *J. Hydrometeor.*, **16**, 579–597, https://doi.org/10.1175/JHM-D-14-0103.1

Ramos A.M., R. Nieto, R. Tomé, L. Gimeno, R.M. Trigo, M.L.R. Liberato and D.A. Lavers (2016) Atmospheric rivers moisture sources from a Lagrangian perspective. Earth Syst. Dynam. 7: 371-384.Ralph, F. M., and M.D. Dettinger (2011), Storms, floods, and the science of atmospheric rivers. *Eos, 92*(32), 265–266. https://doi.org/10.1029/2011EO320001

Ramos, AM, Wilson, AM, DeFlorio, MJ, *et al*. 2018 International Atmospheric Rivers Conference: Multi-disciplinary studies and high-impact applications of atmospheric rivers. *Atmos Sci Lett*. 2019;e935. https://doi.org/10.1002/asl.935

Ralph, F.M., T. Coleman, P.J. Neiman, R.J. Zamora, and M.D. Dettinger, 2013: Observed Impacts of Duration and Seasonality of Atmospheric-River Landfalls on Soil Moisture and

Runoff in Coastal Northern California. *J. Hydrometeor.*, **14**, 443–459, https://doi.org/10.1175/JHM-D-12-076.1

Ralph, F.M., M.D. Dettinger, M.M. Cairns, T.J. Galarneau, and J. Eylander, 2018a: Defining "Atmospheric River": How the Glossary of Meteorology Helped Resolve a Debate.*Bull. Amer. Meteor. Soc.*, **99**, 837–839, https://doi.org/10.1175/BAMS-D-17-0157.1

Ralph, F. M., A. M. Wilson, T. Shulgina, B. Kawzenuk, S. Sellars, J. J. Rutz, M. Asgari-Lamjiri, E. A. Barnes, A. Gershunov, B. Guan, K. Nardi, T. Osborne, and G. A. Wick, 2018b: Comparison of Atmospheric River Detection Tools: How Many Atmospheric Rivers Hit Northern California's Russian River Watershed? https://doi.org/10.1007/s00382-018-4427-5.

Ralph, F.M., J.J. Rutz, J.M. Cordeira, M. Dettinger, M. Anderson, D. Reynolds, L.J. Schick, and C. Smallcomb, 2019: A Scale to Characterize the Strength and Impacts of Atmospheric Rivers. *Bull. Amer. Meteor. Soc.*, **100**, 269–289, https://doi.org/10.1175/BAMS-D-18-0023.1

Rutz, J.J., W.J. Steenburgh, and F.M. Ralph, 2014: Climatological Characteristics of Atmospheric Rivers and Their Inland Penetration over the Western United States. *Mon. Wea. Rev.*, **142**, 905–921, https://doi.org/10.1175/MWR-D-13-00168.1

Rutz, J.J., W.J. Steenburgh, and F.M. Ralph, 2015: The Inland Penetration of Atmospheric Rivers over Western North America: A Lagrangian Analysis. *Mon. Wea. Rev.*,**143**, 1924–1944, https://doi.org/10.1175/MWR-D-14-00288.1

Sellars, S.L., X. Gao, and S. Sorooshian, 2015: An Object-Oriented Approach to Investigate Impacts of Climate Oscillations on Precipitation: A Western United States Case Study. *J. Hydrometeor.*, **16**, 830–842, https://doi.org/10.1175/JHM-D-14-0101.1

Shields, C. A., and J. T. Kiehl (2016a), Atmospheric river landfall-latitude changes in future climate simulations, *Geophys. Res. Lett.*, 43, 8775–8782, doi: 10.1002/2016GL070470.

Shields, C. A., and J. T. Kiehl (2016b), Simulating the Pineapple Express in the half degree Community Climate System Model, CCSM4, Geophys. Res. Lett., 43, 7767–7773, doi:10.1002/2016GL069476.

Shields, C. A., Rutz, J. J., Leung, L.-Y., Ralph, F. M., Wehner, M., Kawzenuk, B., Lora, J. M., McClenny, E., Osborne, T., Payne, A. E., Ullrich, P., Gershunov, A., Goldenson, N., Guan, B., Qian, Y., Ramos, A. M., Sarangi, C., Sellars, S., Gorodetskaya, I., Kashinath, K., Kurlin, V., Mahoney, K., Muszynski, G., Pierce, R., Subramanian, A. C., Tome, R., Waliser, D., Walton, D., Wick, G., Wilson, A., Lavers, D., Prabhat, Collow, A., Krishnan, H., Magnusdottir, G., and Nguyen, P.: Atmospheric River Tracking Method Intercomparison Project (ARTMIP): project goals and experimental design, Geosci. Model Dev., 11, 2455-2474, https://doi.org/10.5194/gmd-11-2455-2018, 2018.

Trenberth, K.E., 1991: Storm Tracks in the Southern Hemisphere. *J. Atmos. Sci.*, **48**, 2159–2178

Viale, M., R. Valenzuela, R. Garreaud, and F.M. Ralph, 2018: Impacts of atmospheric rivers on precipitation over southern South America. Journal of Hydrometeorology, 19, 1671–1687, https://doi.org/10.1175/JHM-D-18-0006.1

Warner, M.D., C.F. Mass, and E.P. Salathé, 2015: Changes in Winter Atmospheric Rivers along the North American West Coast in CMIP5 Climate Models. *J. Hydrometeor.*,**16**, 118–128, https://doi.org/10.1175/JHM-D-14-0080.1

Wick, G.A., P. J. Neiman and F. M. Ralph, 2013: Description and Validation of an Automated Objective Technique for Identification and Characterization of the Integrated Water Vapor Signature of Atmospheric Rivers, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 4, pp. 2166-2176.

Zhu, Y. and R.E. Newell, 1998: A Proposed Algorithm for Moisture Fluxes from Atmospheric Rivers. *Mon. Wea. Rev.*, **126**, 725–735, https://doi.org/10.1175/1520-0493(1998)126<0725:APAFMF>2.0.CO;2

Table 1.

Listed algorithms (A1, A2, etc.) have submitted catalogues for Tier 1. Algorithm numbering is determined by overall alphabetical order with the first set of numbers assigned to developers who participated in the 1-month proof of concept (A1 – A15), (see GMD paper), followed by Tier 1-only participation (A16-24). ARTMIP algorithm identifier will remain consistent across tiers and scientific papers. Shorthand identifiers also are listed in the developer column (italicized in parenthesis) to help navigate figure interpretation.

Developer *(algorithm name)*	**Type**	**Geometry Requirements**	**Threshold Requirements**	**Temporal Requirements**	**Region**	**DOI/ Reference**
Brands et al. (*Brands_v1, Brands_v2, Brands_v3*)	Condition	>1500km	Both Latitude-dependent Relative and Absolute IVT: 95th (A16)/90th (A23)/90th (A24)/percentile at point of detection, 90th /85th/85th percentile along the AR structure, consider all months for threshold calculation with a min threshold of 240kgm^{-1}s^{-1}. /500kgm^{-1}s^{-1}/250kg m^{-1}s^{-1}Spatial tracking guided by vector IVT.	Time slice	150ºW to 30ºE, 30ºN to 62ºN	10.1007/s00382-016-3095-6
Gershunov et al.[b] (*Gershunov*)	Condition and Track	>= 1500km long	Absolute: 250kgm^{-1}s^{-1} IVT	Time stitching -18 hours (3 time	Western U.S.	10.1002/2017GL074175

			1.5cm IWV	steps for 6 hourly data)		
Goldenson[b] (*Goldenson*)	Condition	>2000km long and < 1000km wide, Object recognition	Absolute: 2cm IWV	Time slice	Western U.S.	Goldenson et. al, submitted
Guan and Waliser[bc] (*Guan_Waliser*)	Condition	Length >2000km and length-width ratio>2; Coherent IVT direction within 45^0 of AR shape orientation and with a poleward component	Latitude-dependent Relative: 85^{th} percentile IVT; Absolute min requirement designed for polar locations: $100 kgm^{-1}s^{-1}$ IVT	Time slice	Global	10.1002/2015JD024257; 10.1175/JHM-D-17-0114.1
Hagos et al. [b] (*PNNL1_Hagos*)	Condition	Dependent on threshold requirements to determine footprint; > 2000 km long and < 1000 km wide	Absolute: 2cm IWV $10 ms^{-1}$ wind speed	Time slice	Western U.S.	10.1175/JCLI-D-14-00567.1
Lavers et al. (*Lavers*)	Condition	4.5^o latitude movement allowed	Latitude-independent Relative: ~85^{th} percentile determined by evaluation of reanalysis products	Time slice	UK,	10.1029/2012JD018027

Leung and Qian[b] (*PNNL2_LQ*)	Track	Moisture flux has an eastward or northward component at landfall; tracks originating north of 25N and east of 140W are rejected	Absolute: mean IVT along track > 500 $kgm^{-1}s^{-1}$ and IVT at landfall > 200 $kgm^{-1}s^{-1}$; grid points up to 500km to the north and south along the AR tracks are included as part of the AR if their mean IVT > 300 $kgm^{-1}s^{-1}$	Time slice	Western U.S.	10.1029/2008GL036445
Lora et al.[b] (*Lora_global, Lora_NPac*)	Condition	Length >= 2000km	Latitude-independent Relative: IVT 100$kgm^{-1}s^{-1}$ above climatological area means for N. Pacific	Time slice	Global (A6), North Pacific (A7)	10.1002/2016GL071541
Mundhenk et al. *(Mundhenk)*	Condition	>1400km length, aspect ratio 1:4, lat limit >16N/S, axis orientation based on IVT	Latitude-dependent Relative IVT percentiles	Time slice	Global	10.1175/JCLI-D-15-0655.1
Muszynski et al. *(TDA_ML)*	Condition	Topological analysis and machine learned	Threshold-free	N/A	Western U.S., adaptable to other regions	Experimental
Payne and Magnusdottir[bc] et al. *(Payne)*	Condition	Length > 1200km, landfalling only	Latitude-independentRelative: 85th Percentile of maximum IVT (1000-500mb)	Time stitching (12-hour minimum) -	Western U.S.	10.1002/2015JD023586;

			Absolute: IWV >2cm, 850mb wind speed > 10m/s			10.1002/2016JD025549
Ramos et al.[bc] *(Ramos)*	Condition	Detected for reference meridians, length >=1500km, latitudinal movement <4.5^{0}N	Latitude-dependent Relative: IVT 85th percentile (1000-300mb)	Time slice, but 18-hour minimum for persistent ARs	Western Europe, South Africa, adaptable to other regions	10.5194/esd-7-371-2016
Rutz et al.[b] *(Rutz)*	Condition	Length >= 2000km	Absolute: IVT (surface to 100mb) = 250kgm^{-1}s^{-1}	Time slice	Global, low value on tropics	10.1175/MWR-D-13-00168.1
Sellars et al.[b] *(CONNECT500, CONNECT700)*	Track	Object identification	Absolute: IVT, thresholds tested = 300 (A11), 500 (A12), 700 (A13) kgm^{-1}s^{-1}	Time stitching, minimum 24-hour period	Global	10.1002/2013EO320001; 10.1175/JHM-D-14-0101.1
Shields and Kiehl[b] *(Shields)*	Condition	Ratio 2:1, length to width grid points min 200km length; 850mb wind direction from specified regional quadrants, landfalling only	Latitude-dependent Relative: [a]ZN moisture threshold using IWV; Wind threshold defined by regional 85th percentile	Time slice	Western U.S. Iberian Peninsula, UK, adaptable but regional specific	10.1002/2016GL069476; 10.1002/2016GL070470

			850mb wind magnitudes			
TEMPEST[b] *(TEMPEST)*	Track	Laplacian IVT thresholds most effective for widths >1000km; cluster size minimum = 120000km^2	IVT >=250kgm^{-1}s^{-1};	Time stitching (time slice only for this study)	Global, but latitude >=15^o	Experimental
Walton et al. *(Walton)*	Condition and Track	Length >= 2000 km	Latitude-dependent Relative: IVT > 250 kg/m/s + daily IVT climatology	Time stitching	Western U.S.	Experimental

[a]ZN relative threshold formula: $Q >= Q_{zonal\_mean} + AR_{coeff} (Q_{zonalmax} - Q_{zonamean})$ where Q = moisture variable, either IVT (kg m^{-1}s^{-1}) or IWV(cm). AR_{coeff} = 0.3 except where noted. (Zhu and Newell, 1998). The Gorodetskaya method uses $Q_{sat,\ where}$ Q_{sat} represents maximum moisture holding capacity calculated based on temperature (Clausius-Clapeyron), an important distinction for polar ARs. Additional analysis on the ZN method can be found in Newman et al., 201

[b]Methods used in a 1-month proof-of-concept test (Section 5). These methods are assigned an algorithm id, i.e. A1, A2, etc.

[c] These 1-month proof-of-concept methods apply a percentile approach to determining ARs. A3 and A8 applied the full MERRA2 climatology to compute percentiles. A9, applied the Feb 2017 climatology for this test only. For the full catalogues, A9 will apply extended winter and extended summer climatologies to compute percentiles. Please refer to individual publications (DOI reference column in this table) for climatologies used in earlier published studies by each developer. The climatology used to compute percentile is often dependent on the dataset (re-analysis or model data) being used.

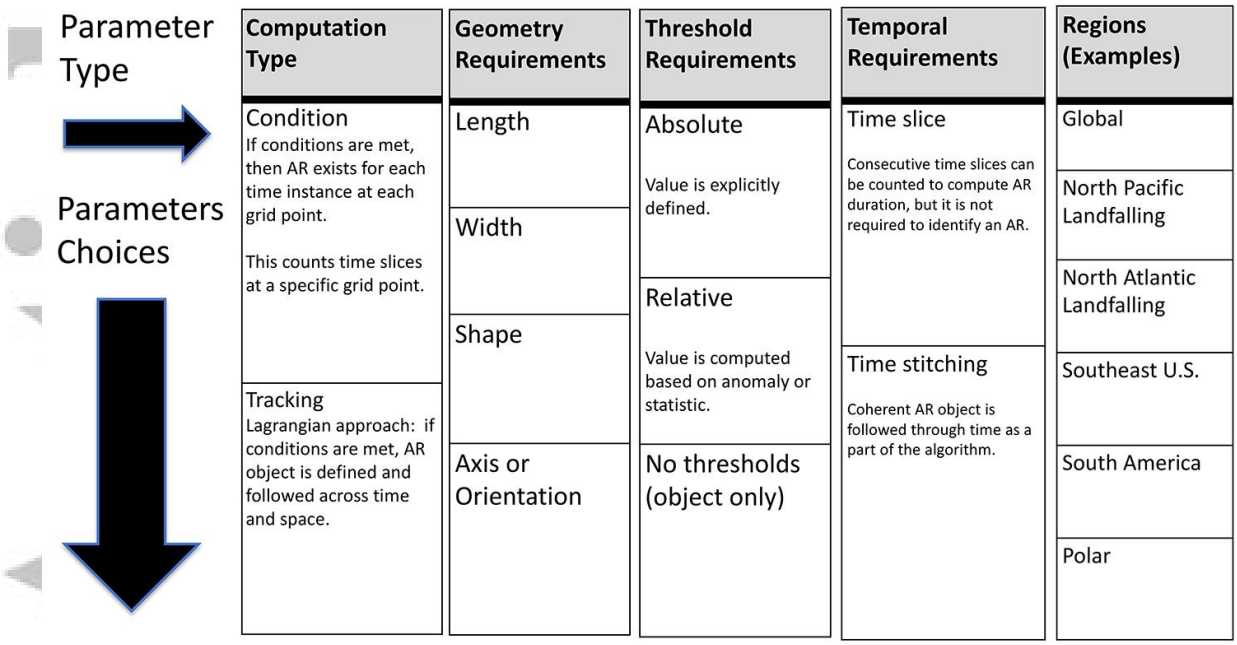

Fig. 1: Schematic diagram illustrating the diversity of AR identification and tracking methods found in current literature by categorizing the variety of parameters used as criteria, and then listing different types of choices available per category.

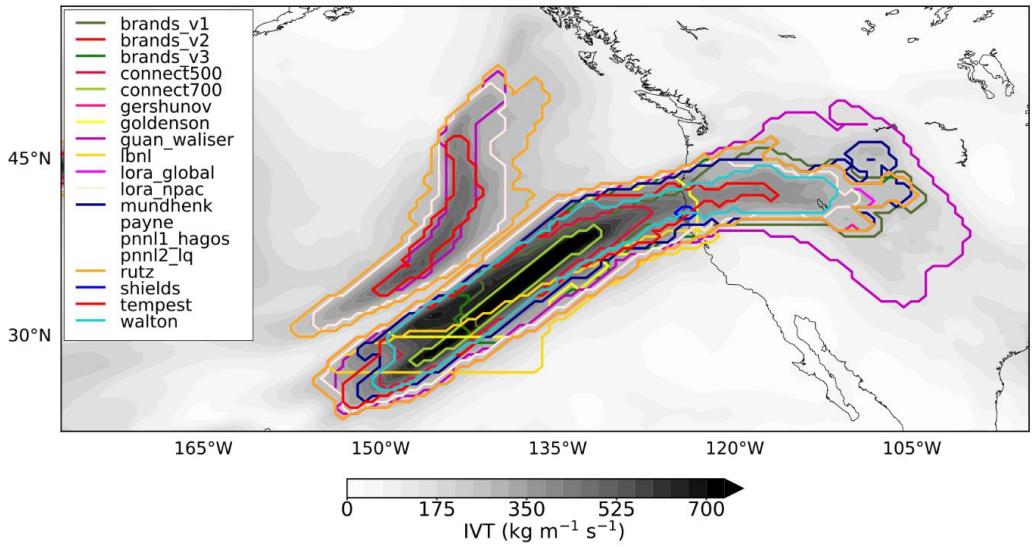

Fig. 2: Example of how AR identification and tracking methods differ over the northeastern Pacific, based on MERRA v2 data from 0000 UTC 15 February 2014. Gray shading represents IVT (kg m^{-1} s^{-1}), and colored contours represent the spatial regions designated as ARs by the various methods. Note that only algorithms available in this region are shown.

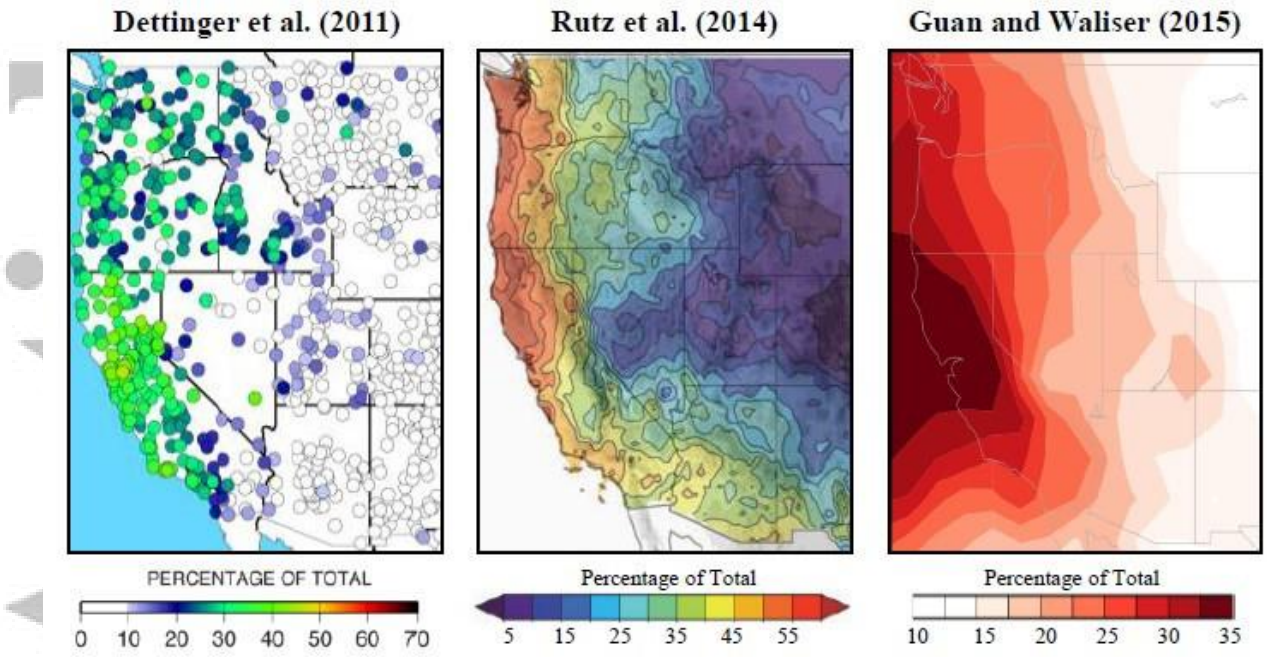

Fig. 3: Fraction of total cool-season precipitation attributable to ARs from (a) Dettinger et al. (2011) and (b) Rutz et al. (2014). (c) As in panels (a, b), but for annual precipitation from Guan and Waliser (2015). These studies use different AR identification methods, as well as different atmospheric reanalyses, observed precipitation data sets, and methods of attributable precipitation to ARs.

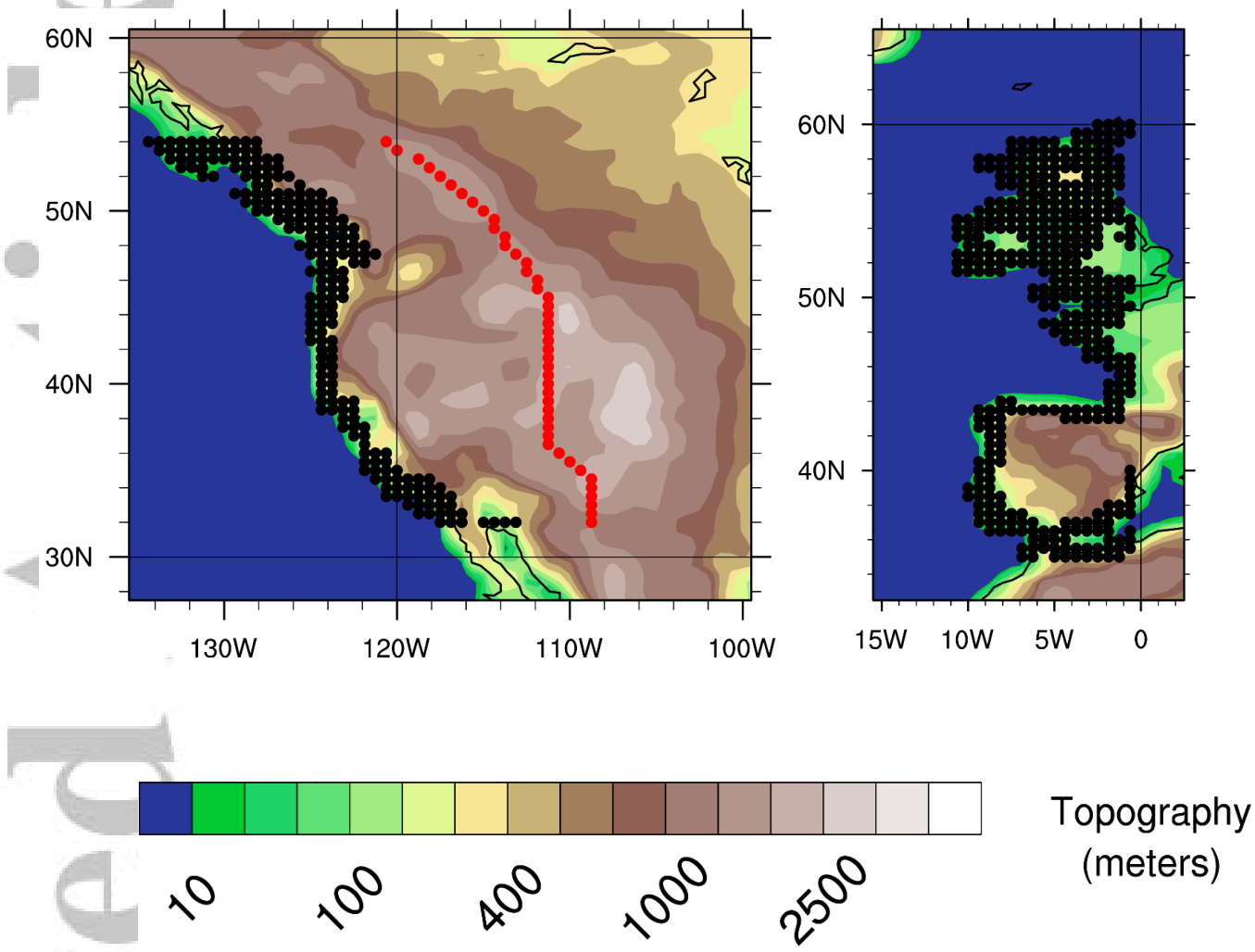

Fig. 4: Selected transects along the North American West Coast (left panel, black dots), through interior western North America (left panel, red dots), and along the European West Coast (right panel, black dots). The coastal transect points are determined by selecting all MERRA-2 reanalysis grid points that have fractional land/sea cover between 32–55°N (and 130–115°W) for North America, and between 35–62°N (and 15°W –10°E) for Europe. The interior transect points are determined by subjectively selecting grid points that represent a significant topographic "crest" between 32–55°N.

Absolute Methods	Relative Methods		
	Latitude-dependent		
	Anomaly above zonal mean	Anomaly above climatology	Percentile based
CONNECT500 *CONNECT700* *Gershunov* *Goldenson* *PNNL1_Hagos* *PNNL2_LQ* *Rutz* *TEMPEST*	*Shields*	*Mundhenk* *Walton*	Brands_v1 *Brands _v2* *Brands_v3* *Guan_Waliser* *Ramos* *Viale*
Machine Learning	Latitude-independent		
TDA_ML	*Lavers* *Payne*		*Lora_global* *Lora_npac*

Global Methods	Regional Methods	
CONNECT500 *CONNECT700* *Guan_Waliser* *Lora_global* *Mundhenk* *Rutz* *Tempest*	*Brands_v1*	*(Europe)*
	Brands_v2	*(Europe)*
	Brands_v3	*(Europe)*
	Gershunov	*(western U.S.)*
	Goldenson	*(western U.S.)*
	Lavers	*(western Europe)*
	Lora_npac	*(North Pacific)*
	Payne	*(western U.S.)*
	PNNL1_Hagos	*(western U.S.)*
	PNNL2_LQ	*(western U.S.)*
	Ramos	*(western Europe, south Africa, adaptable to other regions)*
	Shields	*(western U.S., western Europe, adaptable to other regions)*
	TDA_ML	*(western U.S., adaptable to other regions)*
	Viale	*(southwestern South America)*
	Walton	*(western U.S.)*

Fig. 5: Tables showing the names of ARTMIP Tier 1 methods grouped into (top) absolute / relative / machine learning clusters and (bottom) global / regional clusters. For the bottom table, the region(s) over which data is used from each method are given in parenthesis following the method name. Note that this is not a comprehensive list of all AR identification and tracking methods found in scientifically relevant literature; only those methods used in this study are shown.

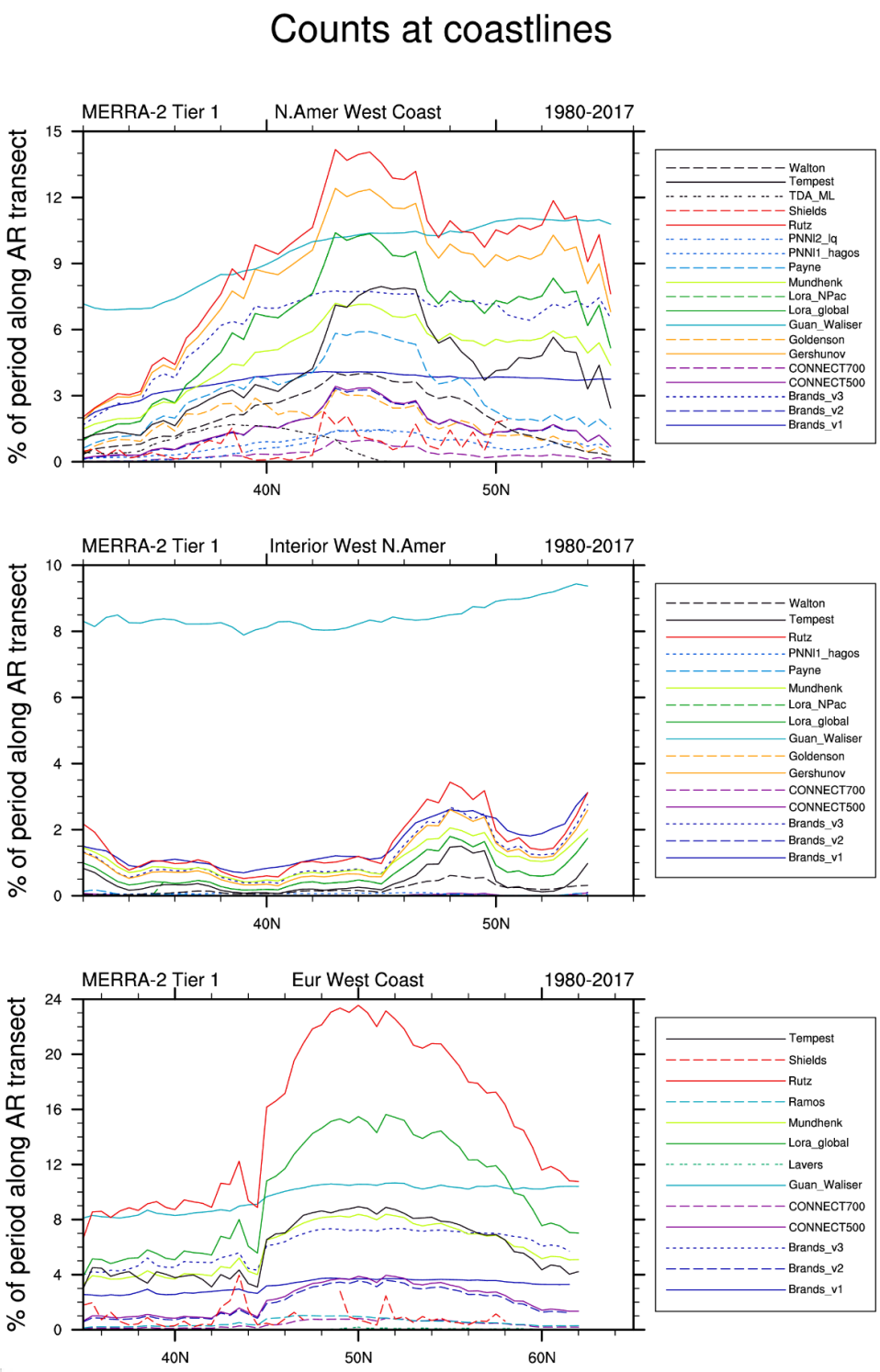

Fig. 6: AR frequency of ARTMIP methods for selected transects (a) along the North American West Coast, (b) through interior western North America, and (c) along the European West Coast. Note that some methods are only available over certain regions.

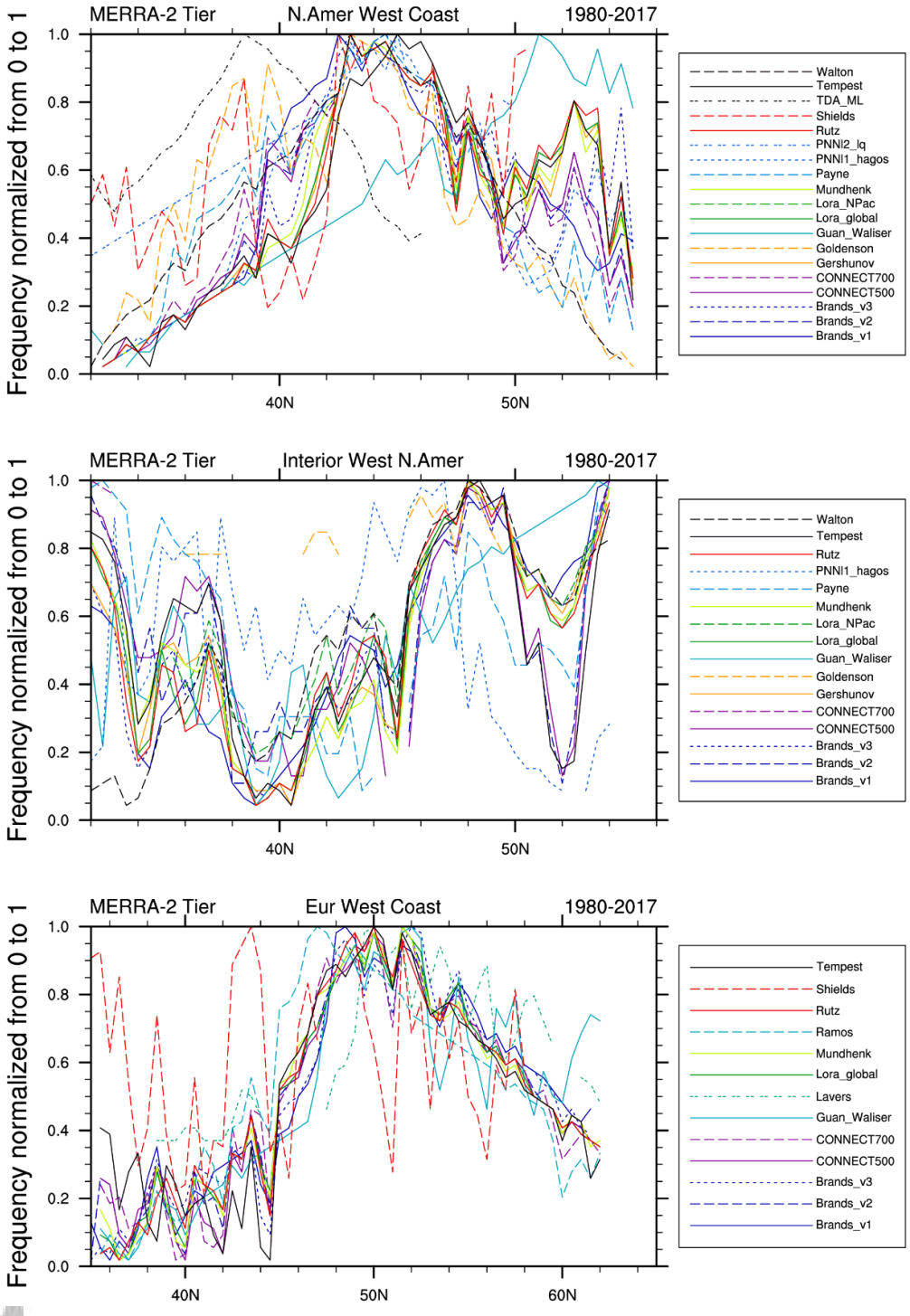

Fig. 7: Normalized AR frequency of ARTMIP methods for selected transects (a) along the North American West Coast, (b) through interior western North America, and (c) along the European West Coast. Note that some methods are only available over certain regions.

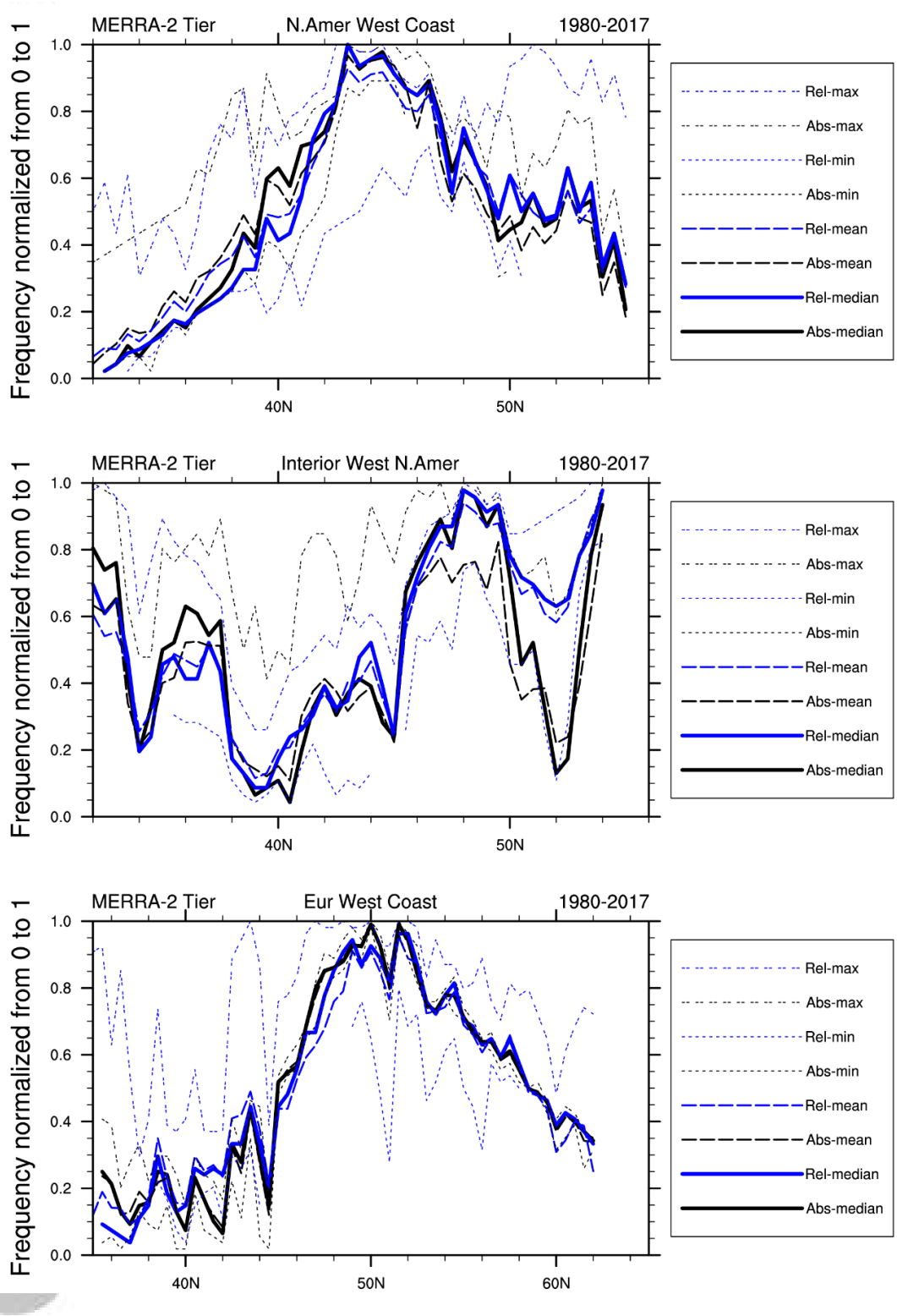

Fig. 8: Normalized and clustered (based on *absolute* or *relative* thresholds) AR frequency of ARTMIP methods for selected transects (a) along the North American West Coast, (b) through interior western North America, and (c) along the European West Coast. Note that some methods are only available over certain regions.

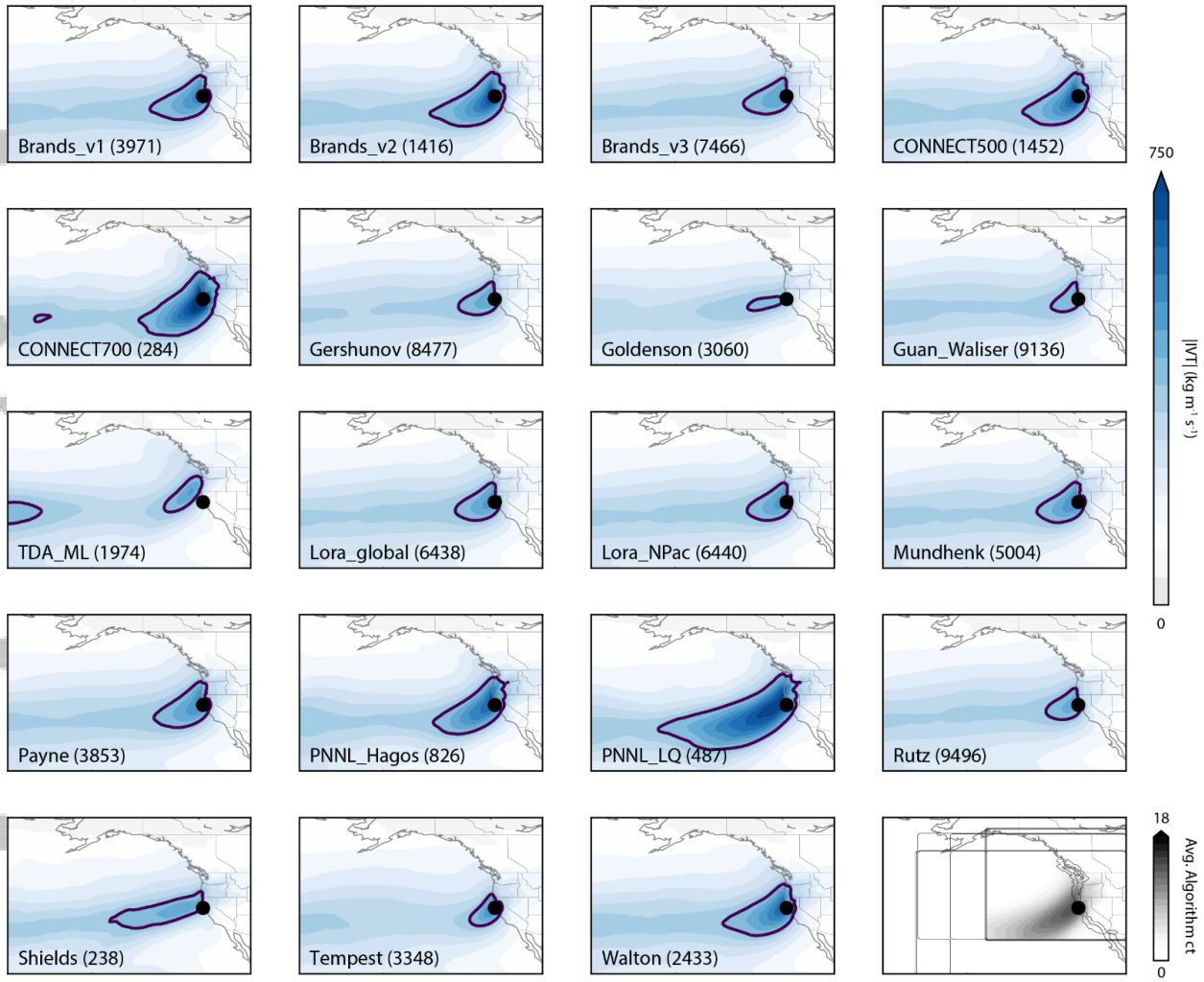

Fig. 9: (all except bottom right) For each method, composites of IVT at all times when that method identifies an AR at the coastal location of 39°N, 123.75°W (black circle). Blue color shading represents IVT magnitude and purple contour indicates IVT of 350 kg m^{-1} s^{-1}. (bottom right) Shading indicates the number of methods identifying an AR anytime the "PNNL_LQ" method identifies an AR at 39°N, 123.75°W.

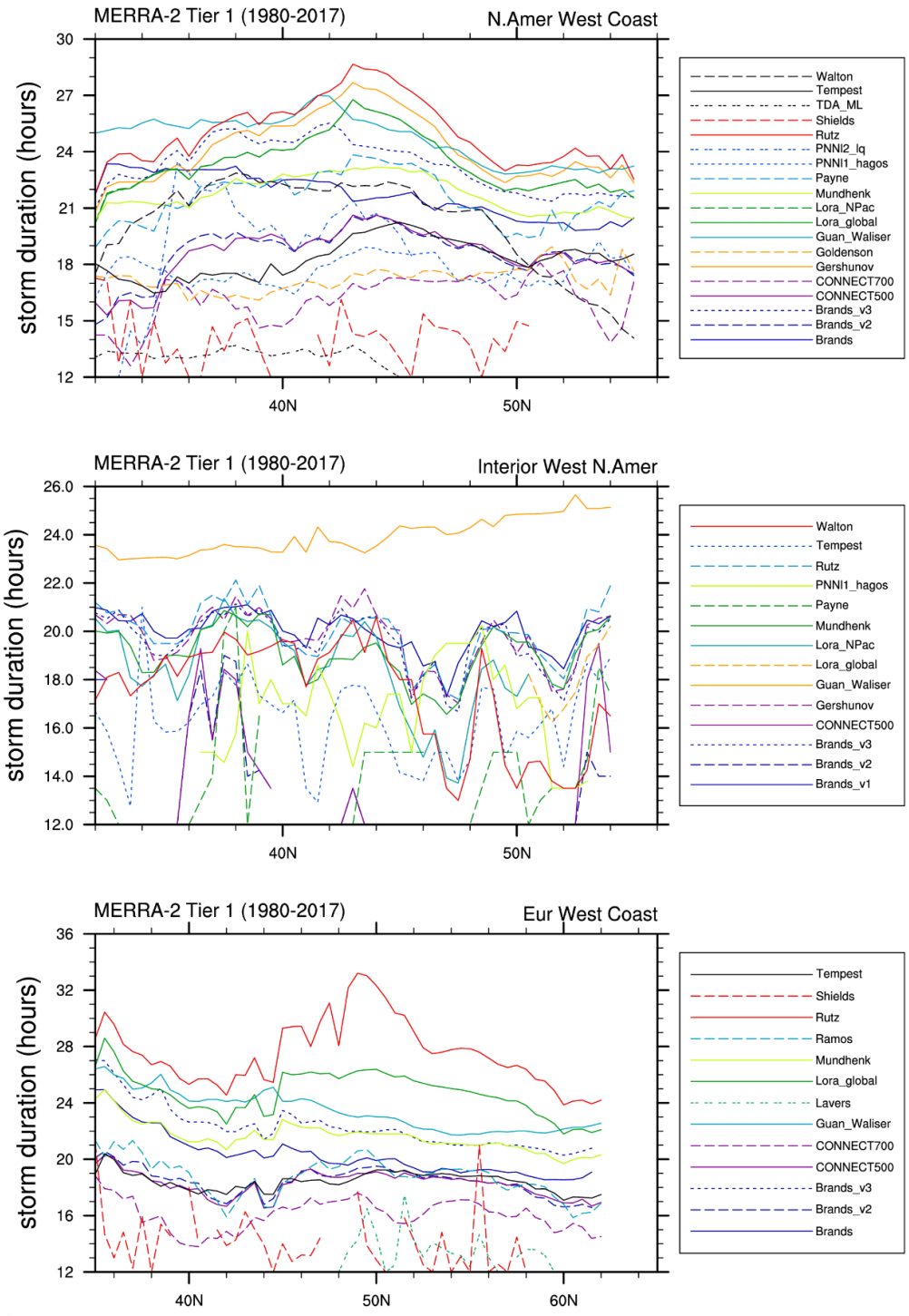

Fig. 10: AR duration of ARTMIP methods for selected transects (a) along the North American West Coast, (b) through interior western North America, and (c) along the European West Coast. Note that some methods are only available over certain regions. Only AR events lasting ≥ 12 h qualify.

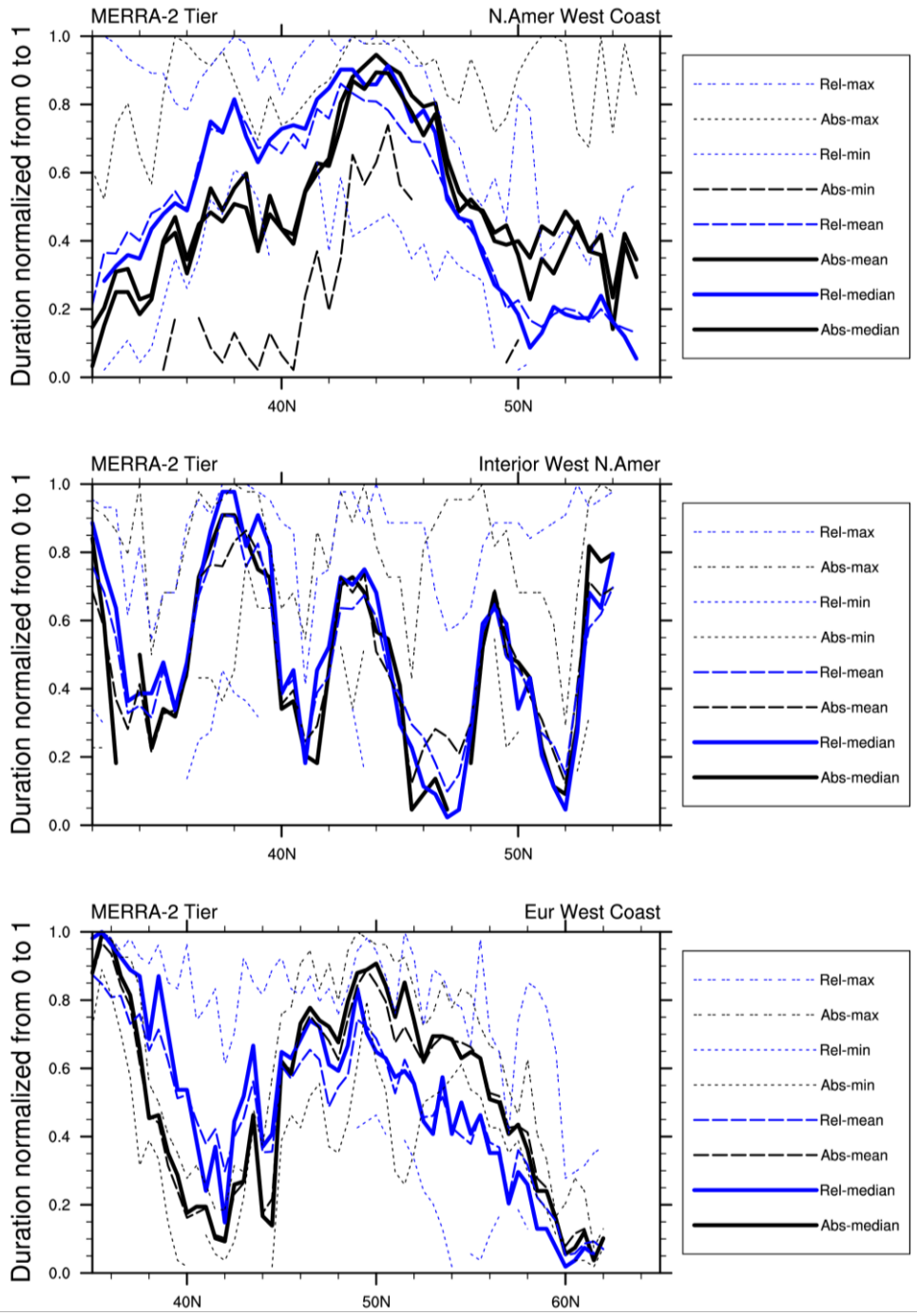

Fig. 11: Normalized and clustered (based on *absolute* or *relative* thresholds) AR duration of ARTMIP methods for selected transects (a) along the North American West Coast, (b) through interior western North America, and (c) along the European West Coast. Note that some methods are only available over certain regions. Only AR events lasting ≥ 12 h qualify.

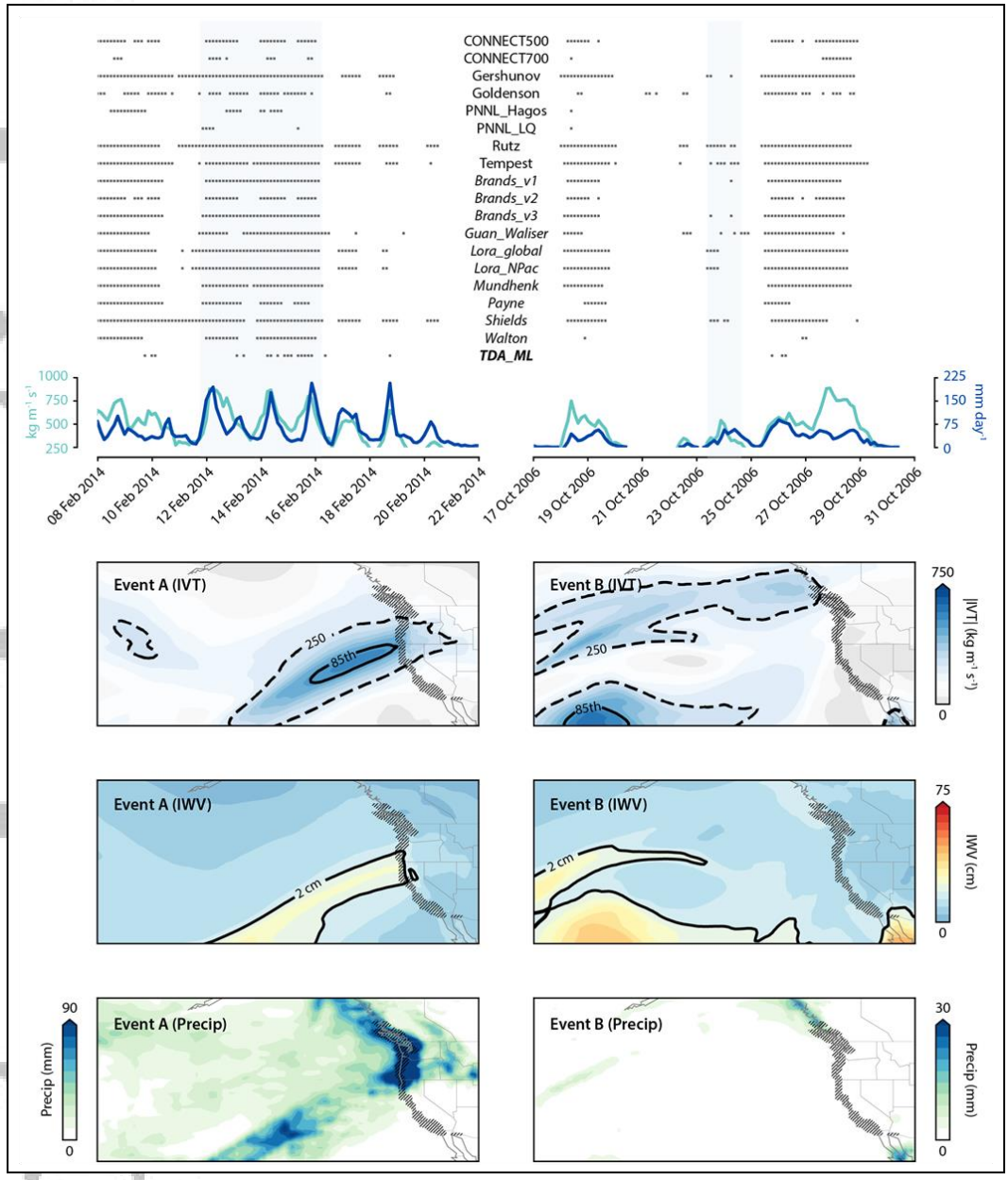

Fig. 12: ARTMIP methods' identification of AR conditions (dots) along a selected transect (hatched in the spatial panels), peak IVT (light blue line, with values below 250 kg m^{-1} s^{-1} dashed) and mean precipitation (blue line) along the transect. Composite of IVT (IVT ≥ 250 kg m^{-1} s^{-1} and IVT ≥ 85th percentile contoured as black dashed and black solid lines, respectively), composite of IWV (IWV ≥ 20 mm contoured as a solid black line), and cumulative precipitation for events centered on (a) 12–16 February 2014 and (b) 23–24 October 2006. The time steps composited for each event are lightly shaded in the top panel. Listed methods use *relative* thresholds if italicized, *no thresholds* if bolded, and *absolute* thresholds otherwise.

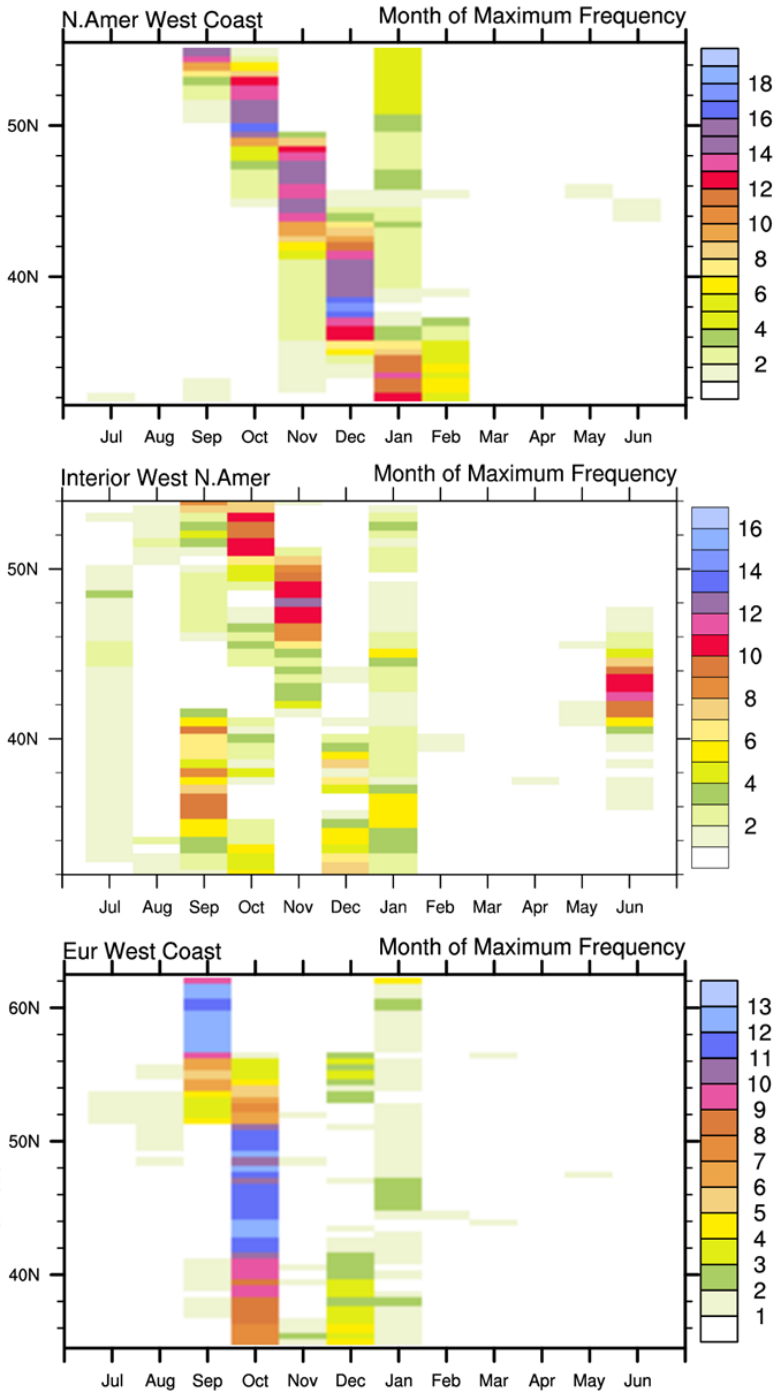

Fig. 13: AR seasonality (month of maximum frequency) of ARTMIP methods for selected transects (a) along the North American West Coast, (b) through interior western North America, and (c) along the European West Coast. Note that some methods are only available over certain regions. Color shading indicates the number of methods for which a given month is the month of maximum AR frequency at each latitude.

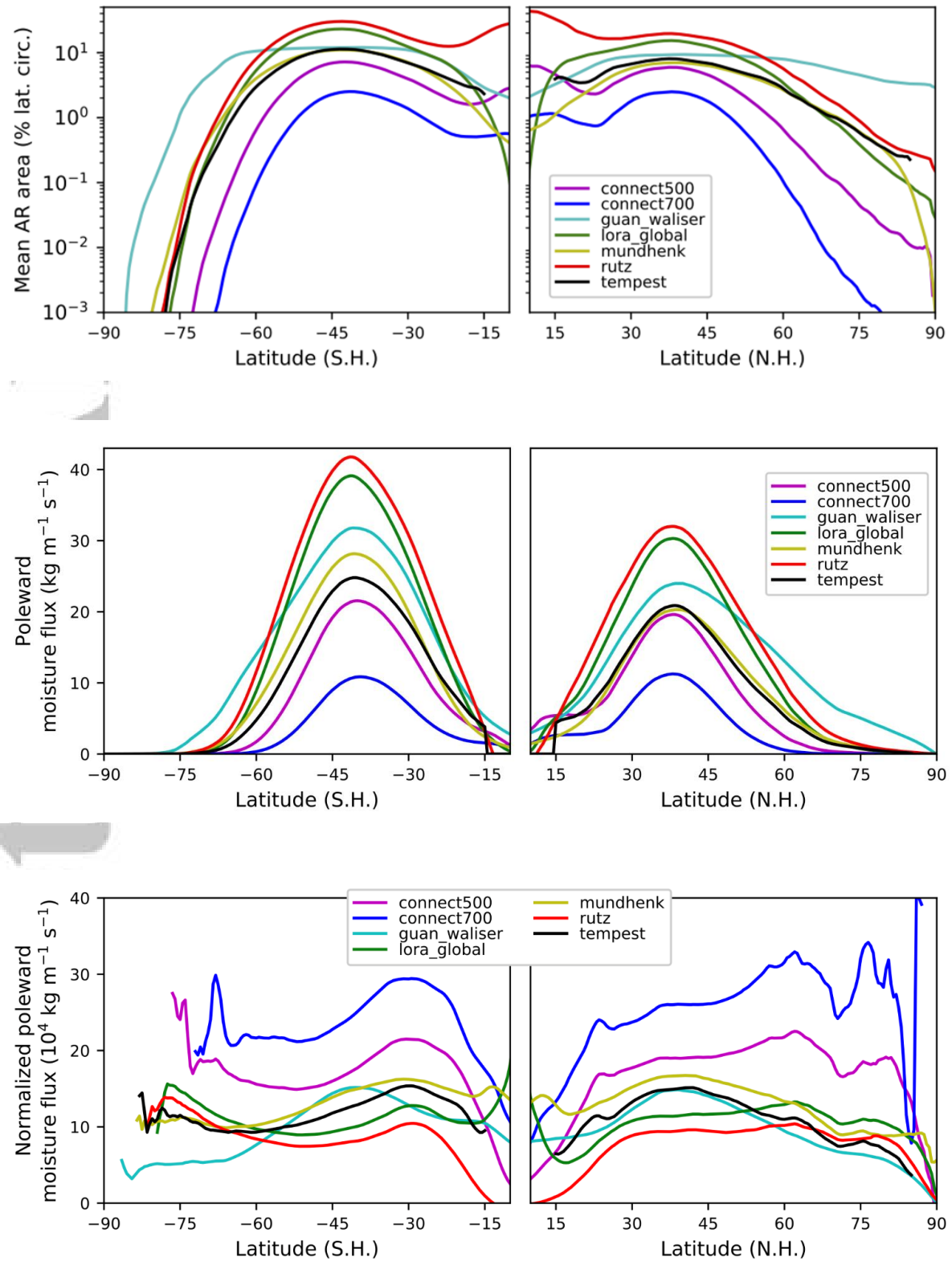

Fig. 14: AR mean area (top), poleward IVT (center), and "efficiency" (bottom) of the ARs identified and tracked by the various ARTMIP methods.

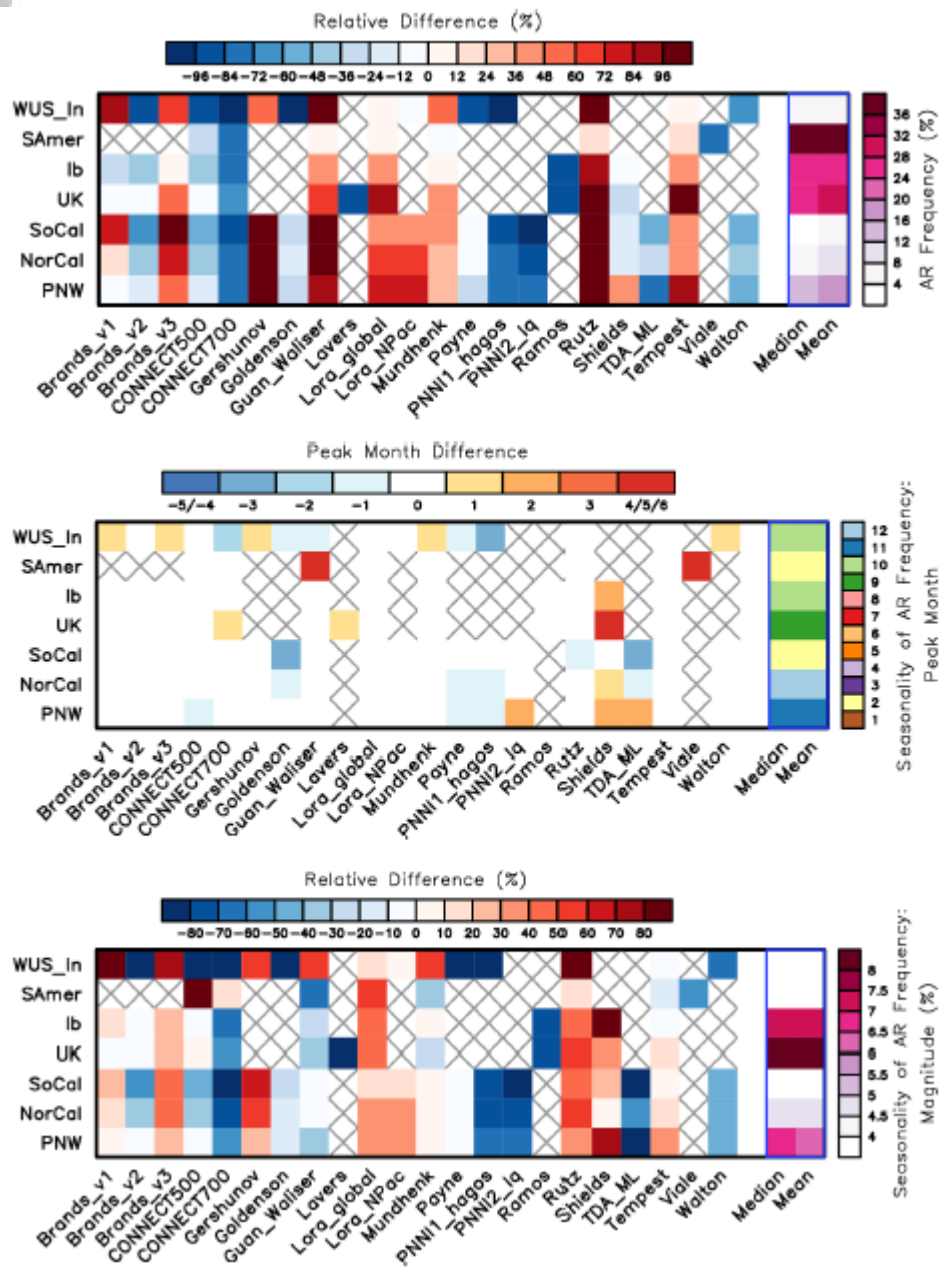

Fig. 15: Diagrams showing the relative difference of results from each ARTMIP method to the all-method median for the metrics of annual (top) AR frequency, (center) month of peak AR frequency, and (bottom) seasonality (or range) of AR frequency. Results are shown for coastal transects of the Pacific Northwest (PNW), northern California (NorCal), southern California (SoCal), the interior western U.S. (WUS_In), South America (SAmer), the United Kingdom (UK), and Iberia (Ib).