Response to Referee #1

June 30, 2020

Many thanks for your time and efforts on this manuscript. I'm giving a one-to-one response to your comments below.

Comment:1. It would be helpful to mention in the introduction that different magnitude thresholds also cause uncertainties in AR measures with reanalysis (as mentioned in ARTMIP publication), not only in future ARs.

Thanks, that is a good point. We are planning to add the following sentences to the end of Line 37:

> For the estimate of present day ARs, different choices of magnitude threshold may also cause considerable uncertainties (Ralph et al. (2018); Shields et al. (2018)). For instance, when estimating the number of landfalling AR events at Bodega Bay using the MERRA2 reanalysis, raising the IVT magnitude threshold from 250 to 500 kg/m/s was found to reduce the total number of AR events during water years of 2005-2016 from 185 (termed as "baseline ARs") to 14 (termed as "stronger ARs"), with the same detection method by Wick et al. 2013a, b (Ralph et al. (2018)).

Comment 2. L 44-45: I don't think this is a disadvantage because some detection methods focus on understanding the landfalling impact of ARs - an absolute threshold will be quite informative and has implications for water management. Depending on scientific questions, the most suitable method changes.

I agree that the most suitable method changes and no silver bullet method exists that suits for all purposes. But L44-45 does not seem to be contradictory to this claim. What it was saying is that the numerical value used as the IWV threshold mostly likely has to be different from that used as the IVT threshold, for instance 20 mm for IWV v.s. 250 kg/m/s for IVT, and we will have to design a suitable choice for IWV and IVT separately. For IWV alone, people found that the 2 cm threshold does not work well for polar climate, so an empirical adjustment method was used. Other than these, using what threshold level is totally up to the specific interest. So perhaps we could phrase the sentences a bit differently, like:

> The prescribed threshold approach also requires different thresholds for IWV-based and IVT-based applications (for instance 2 cm for IWV and 250 kg/m/s for IVT), and in both

cases, possibly different sets of thresholds for mid-latitude systems and polar systems. As demonstrated in (Gorodetskaya et al., 2014), lower air temperature and the reduced water holding capacity demand a separate set of threshold catered to the polar climate.

Having said that, I'd like to share some thoughts on the automated AR detection topic, since this is considered "open discussion" stage. This is not arguing against reviewers comments, but trying to join and inspire the discussion on the subject.

I tend to believe that at least for the problem of "automated AR detection", method IS definition, and there is practically no other definition beyond one's method. Let us assume a continuous spectrum in the AR definition from the extremely implicit end to the extremely explicit end. On the extreme implicit end, one may define ARs as "long and strong water plumes in the lower atmosphere ...", on the explicit end, ARs are whatever someone's code deems as ARs. Obviously the practically useful definition is somewhere in between, but where? Now move one step away from the extreme implicit end and elaborate on the "long" attribute: how long is long? As soon as one puts a number (or two numbers if ones adopts a hysteresis thresholding approach, or even a length range if some sort of fuzzy logic is used) to it, there is an array of technical details that follow: how to measure the length? From where to where? Does that length measurement consider the shape curvature? At how fine a resolution? One can keep on asking such questions, and we have not yet touched upon other attributes besides length. Therefore this implicit-explicit definition spectrum seems extremely skewed, and one tiny step into the explicit side would result in numerous technical considerations, and ultimately, method is definition. (This of cause does not mean that methods share no overlaps, in fact many of them do to a good extent.)

That is perhaps part of the reason for the considerable discrepancies among studies regarding the "definition of ARs". We all seem to agree on the very implicit definition, but in practice that does not matter much (that also depends on the degree of subtlety one's analysis interest is), because method determines what type of object goes into the analyses. I also think that maybe we are not using the term "definition" carefully enough, because in some cases it seems to be a mixture of "definition" and "selection". Say when one applies an IVT threshold of 500 kg/m/s, according to the Ralph et al. 2018 terminology, this detects "stronger ARs" which are distinguished from "baseline ARs". This seems to be going into sub-categories of ARs already when we have not arrived at a good consensus on what ARs are in general. I do not think it meant to say that systems that are weaker than 500 kg/m/s are not ARs, it just ignores them for some specific reasons (maybe not interesting enough for water management). For tropical cyclones, we have strength-based sub-categorical terms. Weaker ones are called depressions, stronger ones storms, and even stronger ones typhoons and hurricanes, and collectively they can be called TCs. But for ARs we do not seem to be as delicate, at least for now. On the other hand, if similar strength-based sub-categories are to be designed for ARs (for instance in Ralph et al. (2019)), then maybe those detection methods that directly threshold the strength need to be more specific about what term to use to name their detections (remember method is definition), for instance, a method that uses a 500 kg/m/s IVT threshold detects "moderate to exceptional" ARs (using the Ralph et al. 2019 terminology and omitting the duration considerations). Sellars et al. (2017) used 750 kg/m/s, and correspondingly what they found are "strong to exceptional" ARs.

I feel that this is to some extent related to the 1st comment, and again, it is totally subject to the user to select a threshold level for a specific research purpose. And if we could be more specific about the definition from a strength variation point of view, maybe the disagreement on AR definition is not as severe as it appears, in many cases they are all ARs, just in different strength categories.

Comment 3. Figure 3: it would be great to mark latitudes and longitudes.

The figure has been recreated with latitude and longitude axes added. Also added an inset plot showing the IVT distribution and the relevant AR. Please checkout the included figure file.

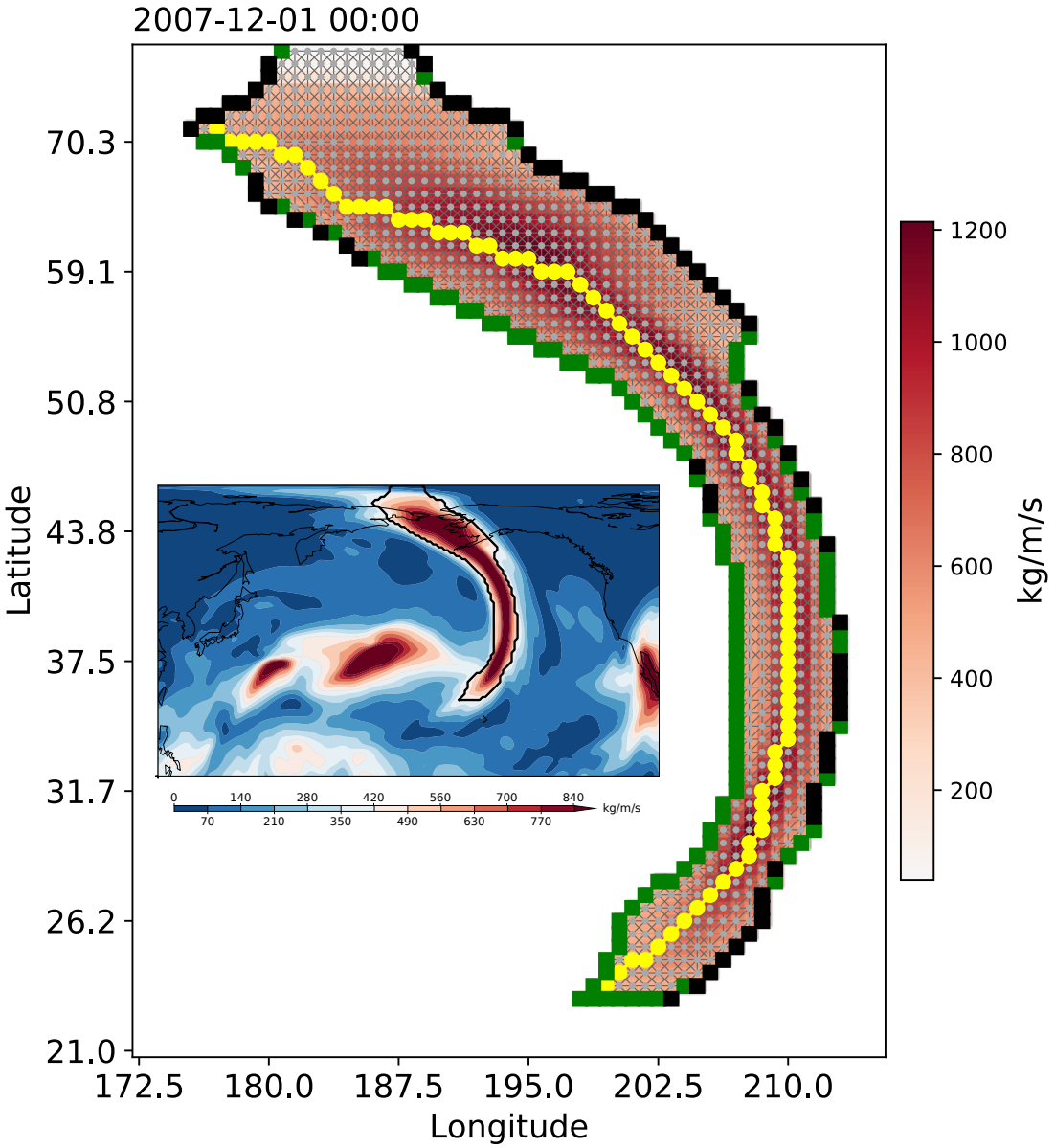

Figure 1: Application of the axis finding algorithm on the AR in the North Pacific, 2007-Dec-1 00 UTC. IVT within the AR is shown as colors, in $kg/m/s$. The region of the AR (I_k) is shown as a collection of gray dots, which constitute nodes of the directed graph. Edges among neighboring nodes are created. A square marker is drawn at each boundary node, and is filled with green if the boundary node has net input moisture fluxes ($n_i \in L_{k,in}$), and black if it has net output moisture fluxes ($n_i \in L_{k,out}$). The found axis is highlighted in yellow. The inset image shows the IVT distribution over North Pacific with the selected AR highlighted in black contour.

Comment 4. L245: Are the tropical reservoirs removed by the algorithm through geometric filtering or they are counted as AR candidates?

These non-zero systems in tropical reservoirs will be removed mostly by the latitudinal requirement that the centroid being North of 23 degree (in the case of Northern Hemisphere), which is listed in the point 4 in the Geometric considerations Section (Line 170). So technically they are counted as AR candidates at the initial selection stage, and are removed by the geometrical filtering.

In implementation, we put the latitudinal filtering as the 1st step, because it is easiest and cheapest to perform, and removing these as early as possible can save some unnecessary computations in other geometrical filtering processes like minimal length or length/width ratio.

Comment 5. Figure 9: what does inf duration mean?

The "inf" label was put there to form the right bin edge for the last bin in the histogram, so that the last bin includes all tracks lasting longer than 150 hours. The bin bars and bin edge labels in the plot have been aligned such that each bin falls in between 2 edges on the left and the right, and "inf" is just the last right edge.

We are going to change the figure caption to:

> Distribution of track durations (in hours) of AR tracks in the North Pacific (cyan) and North Atlantic (orange). The "inf" label is used to form the right bin edge for the last bin which includes all track lasting longer than 150 hours.

to help clarify this.

Minor:1. L51: “filamentary”?

Changed to "filamentary".

References

F. Martin Ralph, Anna M. Wilson, Tamara Shulgina, Brian Kawzenuk, Scott Sellars, Jonathan J. Rutz, Maryam A. Lamjiri, Elizabeth A. Barnes, Alexander Gershunov, Bin Guan, Kyle M. Nardi, Tashiana Osborne, and Gary A. Wick. Artmip-early start comparison of atmospheric river detection tools: how many atmospheric rivers hit northern california's russian river watershed? *Climate Dynamics*, 52:4973–4994, 2018. ISSN 0930-75751432-0894. doi: 10.1007/s00382-018-4427-5.

F. Martin Ralph, Jonathan J. Rutz, Jason M. Cordeira, Michael Dettinger, Michael Anderson, David Reynolds, Lawrence J. Schick, and Chris Smallcomb. A scale to characterize the strength and impacts of atmospheric rivers. *Bulletin of the American Meteorological Society*, 100:269–289, 2019. ISSN 0003-00071520-0477. doi: 10.1175/bams-d-18-0023.1.

S. L. Sellars, B. Kawzenuk, P. Nguyen, F. M. Ralph, and S. Sorooshian. Genesis, pathways, and terminations of intense global water vapor transport in association with large-scale climate patterns. *Geophysical Research Letters*, 44:12,465–12,475, 2017. ISSN 0094-8276. doi: 10.1002/2017gl075495.

Christine A. Shields, Jonathan J. Rutz, Lai-Yung Leung, F. Martin Ralph, Michael Wehner, Brian Kawzenuk, Juan M. Lora, Elizabeth McClenny, Tashiana Osborne, Ashley E. Payne, Paul Ullrich, Alexander Gershunov, Naomi Goldenson, Bin Guan, Yun Qian, Alexandre M. Ramos, Chandan Sarangi, Scott Sellars, Irina Gorodetskaya, Karthik Kashinath, Vitaliy Kurlin, Kelly Mahoney, Grzegorz Muszynski, Roger Pierce, Aneesh C. Subramanian, Ricardo Tome, Duane Waliser, Daniel Walton, Gary Wick, Anna Wilson, David Lavers, Allison Collow, Harinarayan Krishnan, Gudrun Magnusdottir, and Phu Nguyen. Atmospheric river tracking method intercomparison project (artmip): project goals and experimental design. *Geoscientific Model Development*, 11: 2455–2474, 2018. ISSN 1991-9603. doi: 10.5194/gmd-11-2455-2018.