I find this manuscript to be well-written overall, and to describe a feature that seems very desirable as a long-time user of CESM and E3SM. There are a few comments/questions I had though:

1. I found this behavior described in the paper to be odd:

"If the metric and the QoI are both 3D but have different numbers of vertical layers (e.g., the metric is the air temperature defined at layer midpoints while the QoI is the net longwave radiative flux defined at layer interfaces), then masking will be skipped, meaning this specific QoI will be captured for output as if no conditional sampling had happened."

My first reaction to this was that the code should raise an error in this case rather than proceed without conditional sampling, since it's surely a user error if a QoI is listed that is incompatible with the metric dimensions. It's only later that the manuscript mentions that the QoIs have to be the same for all conditions, i.e. there may be multiple conditions with different metrics and the list of QoIs cannot be tailored to each condition separately. I assume that this is the reason why the code has to allow a QoI to be specified in this way rather than raising an error, but when reading the paper in order, it's not clear why this behavior was chosen.

2. In order to better understand how portable the code is, is it true that the conditional_diag module does not use EAM-specific data structures, but that the other three modules do to some extent? Also, since the vertical coordinate must be known in order to perform averaging, is the vertical dimension always the last array dimension in the code?

3. For Tables B1 and B2, it may be good to mention that these order of checkpoints in the table is the same as the actual order of the checkpoints in the code (assuming that this is the case).