We thank all reviewers for their helpful comments. Our responses to each reviewer are below.

Reviewer 1. We thank the reviewer for the positive comments. We will add a figure explaining the full architecture in the camera-ready version of the paper.

Reviewer 2. The reviewer has three major critiques of the paper, which we address in order.

- **Hazard Rate**. The reviewer states that the hazard rate is assumed to be known. This is a misunderstanding. While the hazard parameter can be specified as a hyperparameter if known, it can also be chosen to be a learnable parameter of the model, since the BOCPD procedure is fully differentiable. Indeed, we use a learnable hazard rate on the NBA experiment, as there is no ground truth hazard rate. In all experiments, making the hazard parameter learnable yielded performance comparable to when it was pre-specified as a hyperparameter. We will add further discussion of this to the body of the paper.
- **Problem Setting**. As noted by reviewers 1 and 4, we believe the problem setting presented in the paper extends the domain of applicability of meta-learning tools beyond the standard setting. To showcase the broad applicability of MOCA, we included a diverse array of experiments including on real-world time series data in the NBA example, and as part of a decision making pipeline in the contextual bandit example. The reviewer's concerns center on the prevalence of settings with (1) discrete unlabeled switches in task, (2) exponential task length distributions, and (3) enough task similarity for meta-learning to be effective.
    1. Discrete switches in context are prevalent in diverse settings such as time series forecasting, mobile robotics, and anomaly detection — previous work in all of these settings has leveraged changepoint detection; MOCA leverages changepoint detection to allow meta-learning to be applied to such problems.
    2. We presented MOCA using an exponential distribution on task-length because it is simple (memoryless, single parameter) and broadly applicable. However, the changepoint detection algorithm can operate with a wide range of probability distributions for the task length [Adams & MacKay, 2007], and can be adjusted for a given application if needed.
    3. In principle, meta-learning will be useful as long as there is some commonality between the tasks, and settings in which there is absolutely no similarity between sequential tasks are rare. In robotics, for example, input data comes from the real world and is constrained by the laws of physics; meta-learning would capture these physical priors from data.
- **Point Estimate Changepoint Detection and Computational Efficiency**. MOCA relies on the full run length belief to backpropagate through the changepoint detection algorithm to train the underlying model. While "hard" changepoint detection algorithms exist, backpropagating through these would result in the standard difficulties of backpropagation through samples from a discrete distribution. At test time, one may be able to run a point estimate version of the MOCA algorithm, but the performance of this is likely highly problem-dependent.

Reviewer 4. We clarify that throughout the experiments, the baseline comparisons (TOE, sliding window, etc.) are trained using the same log likelihood loss function as MOCA, and so MOCA does not have a unique bias.

Reviewer 6. As discussed in the response to reviewer 2 ("Problem Setting"), we believe the MOCA problem setting is representative of many real world scenarios. Below, we address the other major comments from the reviewer (approximately) in order.

- **Number of Tasks and Task Recurrence**. MOCA does not directly estimate the number of tasks or changepoints in a sequence. Instead, MOCA reasons only about how much past data to use in making predictions, by maintaining a belief distribution on the run length of the current task. If the task switches to one previously seen in the sequence, MOCA would recognize that the task has changed and start adapting to the new task; MOCA would not recognize that the new task matches a previous one.
- **Label Shift**. Label changes can be handled by MOCA; this is shown for categorical outputs in our classification experiments and continuous outputs in our regression experiments. Indeed, we emphasize that MOCA can leverage changes in both the dependent and independent variable to detect changepoints.
- **IID Data Within Tasks**. We agree that this work assumes iid data generation within each task. In practice however (as demonstrated by the NBA experiment) MOCA can perform well even when this assumption is violated. Indeed, we note that many datasets have some aspect of time dependency that is ignored in the modeling. In any case, we believe our assumptions on the problem setting are stated clearly in section 2.
- **Negative Transfer**. Negative transfer is reduced because MOCA down-weights run length hypotheses that suffer negative transfer and make poor predictions. This is demonstrated via comparison to sliding window models, which do not monotonically improve with longer run lengths. We acknowledge that our statement on "avoiding" negative transfer was likely too strong; we will amend this to say that negative transfer is mitigated instead.
- **TOE with Access to Sequence Position**. The problem setting we consider is invariant under time shifts, so adding sequence position as an input to a model does not provide useful information for making predictions. Therefore, we do not consider this in our experiments.