We thank the reviewers for their insightful feedback. In the following, we address their concerns and questions.

[R1] "**Lacking of novelty... All three modules are known techniques with little technical improvements**": As will be discussed in more detail in the sequel, there might be some significant misunderstanding by the reviewer on modules 2 and 3 (i.e., module 2 is not based on attention; module 3 is not standard deep SVDD loss). Moreover, module 1 is not our main novelty, though for timeseries it is essential to use some model to extract multiscale features. Our main novelty is on developing a hierarchical one-class model for timeseries, which allows diverse patterns of multiscale features from normal timeseries to be captured.

"**Features fusion module using attention mechanism ... section 3.1.2 should cite attention mechanism rather than re-invent some technical terms.**" : There is a big misunderstanding. The proposed method is more similar to clustering than attention, both of which are based on the notion of similarity. If equation (5) is viewed as attention as suggested, we would have $\bar{\mathbf{f}}_t^{l-1}$ as the query, and $\{\mathbf{c}_j^l\}_{j=1}^{K^l}$ as the keys (as normalization is w.r.t. the K^l $\mathbf{c}_j^l$'s), and equation (7) would correspond to the attention output. However, note that the summation in (7) involves K^{l-1} (instead of K^l) terms, and so obviously this cannot be attention.

	MSL (%)			SMAP (%)		
	prec	rec	F1	prec	rec	F1
attention	90.66	87.21	88.90	79.70	66.35	72.41
self-attention	82.49	93.33	87.77	94.55	56.05	70.38
proposed	92.85	94.51	**93.67**	91.33	99.36	**95.18**

Indeed, as mentioned at line 116, this should be understood as a *hierarchical clustering* procedure. The $\bar{\mathbf{f}}_t^{l-1}$'s are clustered by their similarities w.r.t. (learnable) centers $\mathbf{c}_j^l$'s using (5), to form the higher-level $\hat{\mathbf{f}}_{t,i}^l$'s in (7). At the intermediate layers, $\hat{\mathbf{f}}_{t,j}^l$ is further merged with the multiscale feature $\mathbf{f}_t^{l+1}$ (extracted by dilated RNN) to form $\bar{\mathbf{f}}_{t,j}^l$.

For experiments, we add two baselines that use attention to obtain $\hat{\mathbf{f}}_{t,i}^l$ (and uses a one-layer MLP to fuse the query and learned attention vector):[1] (i) attention, using query $\bar{\mathbf{f}}_t^{l-1}$, keys $\{\mathbf{c}_j^l\}_j$, and values $\{\mathbf{c}_j^l\}_j$; (ii) self-attention, using query $\bar{\mathbf{f}}_t^{l-1}$, keys $\{\bar{\mathbf{f}}_\tau^{l-1}\}_{\tau<t}$, and values $\{\bar{\mathbf{f}}_\tau^{l-1}\}_{\tau<t}$. Because of lack of time, experiments are only run on the MSL and SMAP data sets. The table above shows that the proposed model performs best. The discussion above and experimental results will be added to the final version.

"**Anomaly loss using SVDD**": Deep SVDD uses only one center and one layer, while we have multiple centers ($\mathbf{c}_j^L$'s) and multiple layers. Besides the simple extension that sums over all $\mathbf{c}_j^L$'s, the key challenge is on what contribution of each $\bar{\mathbf{f}}_{t,j,s}^L$ at the last layer should be compared with each $\mathbf{c}_j^L$. We propose to aggregate the contributions from the features $\bar{\mathbf{f}}_{t,j,s}^l$'s at all layers via an efficient recursive computation of $\{\tilde{R}_{t,j}^l\}$.

"**Authors do not mention how features extracted from dilated RNNs have been inputed into fusion module**": The reviewer might have overlooked parts of the paper. Indeed, these have been discussed at lines 115-116 and 126-128.

"**Not clear why dilated RNN is a valid choice**": As suggested by the reviewer, the dilated RNN can be replaced by any other model that can extract multiscale features. We used the dilated RNN only as an example model. As discussed above, this part is not our main novelty. Our key novelty is on how to model the diverse patterns of multiscale features extracted from timeseries (by the clustering and anomaly detection modules, and the combination of losses).

[R3] "**Broader Impact... use of multiple indices... Figure 1 (right) is not clear**": Thanks for your suggestion. We will improve our description on one-class learners, simplify notations and make the graph clearer in the final version.

[R4] "**In each layer, there are K^l clusters... motivation**": In SVDD, the ball is used to enclose most of the normal patterns. However, as mentioned in line 66-67, real-world timeseries data may have complex characteristics and so multiple balls can better model the diverse normal temporal behaviors at each resolution (this is similar to using the more flexible Gaussian mixture over a single Gaussian in other machine learning models).

"**So many hyper-parameters (K^l) ... sensitivity regarding K^l**": Important hyper-parameters in our network mainly are λ_{orth} and λ_{TSS}. For K^l, to construct a hierarchical clustering structure, we empirically found that simply using a constant or a decreasing setting such as (6, 6, 6) or (18, 12, 6) for $(K^1, \ldots, K^L)$ can achieve a good performance. Thus, we use a relative small search space to find a good (suboptimal) hyperparameter setting for K^l.

"**Self-supervised and orthogonal losses**": These two losses are important. Without the orthogonal loss, the centers may be very similar or even duplicate; without the self-supervised loss (which is based on timeseries prediction), the model may fail to capture temporal dependencies, which are essential for a proper representation of timeseries data.

"**Theoretical analysis about MVDD**": As in deepSVDD, when we use the soft-boundary SVDD objective in each hypersphere, the proof of their Proposition 4 can be easily extended to our model, and the ν-property holds. Here, we give a brief proof sketch: Define $d_{i,j} = \|\text{NN}_j(x_i; \mathcal{W}_j) - \mathbf{c}_j\|^2$ for $i = 1, 2 \ldots, N$ and $j = 1, 2 \ldots, K^L$. The number of outliers for the jth hypersphere is $N_{\text{out}}^j = |\{i : d_{i,j} > r^2\}|$. It can be shown that the soft-boundary objective of our model can be rewritten as $\left(1 - \frac{\sum_{j=1}^{K^L} R_j^L N_{\text{out}}^j}{\nu N}\right) r^2$. The remaining steps are similar to that in the proof of deepSVDD.

[1] As in the literature on attention models, we will specify the query, keys, and values for each baseline.