**Empirical findings in section 2 are very interesting. Briefly mention them in the introduction.**

We will add a sentence to the introduction emphasizing the significance of our findings in section 2.

**Load conditions used for benchmarking (1 req/min) is too small -- How would the results change, if the actual load was 30 req/min?**

We assume that the time series of cloud SDK benchmarking data is ergodic (i.e. stationary over a long period). Under that assumption, QBETS is insensitive to the measurement frequency (at the cloud SDK monitor) so the correctness and tightness should not be affected by a higher measurement frequency. However, an increase in the frequency by a factor of 30 might foreshorten the duration over which the SLAs are durable. Examining the data shows that the predicted percentile values change most dramatically as a result of change points, and a higher frequency would not change the time between change points (since the changes are likely exogenous). Thus we believe that the results are valid for a higher frequency measurement series.

As for measuring the actual response time of the APIs under a higher load, we assume that the cloud platform is scalable thereby distributing the API load across many physical nodes. This is true for most real world cloud platforms, including Google App Engine and AppScale. The distribution of API load (sometimes coupled with Autoscaling features of the cloud platform), helps maintain API response time steady regardless of the load conditions.

We will address the comment by adding these analytical arguments to the paper.

**Can we show that the system can recognize highly unpredictable code (maybe a loop based on a randomly generated value?) and loosen its performance estimates appropriately?**

We do not know the relationship between code predictability and the prediction results -- only that the codes we examined for GAE do not exhibit this level of random execution behavior. We will address this comment by pointing out a possible investigation of this relationship as part of our future work.

**The main contribution of the paper (QBETS) has been published before**

It is true that QBETS has been published before but it has not been applied in this context. We believe we have made clear both the origin of QBETS and also the contribution the paper makes to cloud computing. Thus we offer no additional clarification of this point.

**For private cloud offerings, where operators may have no limitation of how long a service can run, we may see different execution paths appearing in the applications.**

Cerebro takes an exhaustive approach when making SLA predictions, where it considers "all" the branches in the code. It makes separate predictions for each possible branch, and picks the one with the highest value as the final SLA. Therefore, Cerebro already copes with applications with many execution paths. The Stocks application used in our empirical evaluations puts this feature to the test. We will make this fact clear in our writing.

**Currently Cerebro benchmarks SDK operations on some synthetic datasets. It may make sense to modify the applications so that themselves are able to collect information on how the operations are running on their own data.**

We believe that the use of application instrumentation would likely improve the quality of the results or, at worst, leave them unchanged. One goal of the work is to determine the degree to which a non-invasive approach (i.e. one that does not require modification of the application for instrumentation purposes) can be made successful. To address this comment, we will add an investigation of the effects (likely positive) of application instrumentation to future work.

**What happens if there is a lot of variance in the behavior of the system (no stable behavior at all)?**

QBETS requires a minimum number of consecutive measurements (determined from the confidence bound specified and the quantile of interest) to make a prediction. Call this the "measurement set." The variance of the measurement set can be high and QBETS predictions will be correct. However the series must be approximately stationary (i.e. ergodic) meaning that the mean and the variance cannot change abruptly. If they do, QBETS attempts to detect the change and adapt, but the speed with which is does so is bounded by the minimum measurement set required. Thus if changes due to non-stationarity occur more frequently than a complete measurement set can be gathered, QBETS prediction quality will suffer. This degradation will be reflected in the correctness and tightness measures of prediction quality. To address this issue we will attempt to add a clarification to the paper but also to point the interested reader to the relevant exposition of QBETS in the literature.

**How to decide the duration of history data?**

Given the percentile (p) and confidence (c), a theoretical lower bound for the history length can be calculated easily. For c = 0.01 and p = 0.95 this lower bound is around 90 (i.e. 90 data points in the input time series). There is no upper bound for history length, but it makes sense to keep a history of several thousand data points (i.e. a few days of monitoring data). We will briefly mention this in the paper.

**How are the 35 applications chosen?**

We crawled GitHub for Google App Engine apps, and picked a set that were well documented, builds and runs without errors. We will make a note of this in the paper.

**Figures 2 and 3 have a lot of extra space. The y-axis for each can start from 0.6 and 0.4 respectively.**

Will make the changes appropriately.

**In section 3.2 how are the data sizes decided?**

These were determined based on the type of experiments we were planning to carry out using the prototype. However, out system is easily extensible to add benchmarks with other data sizes if necessary (e.g. 10000 or 1000000 entities). We will mention this fact in the paper.