Preliminary Comments from Dr. Steven Hanna on the
03-29-18 Draft CASAC Review of the EPA's *Risk and Exposure Assessment for the Review of the Primary National Ambient Air Quality Standard for Sulfur Oxides*
(External Review Draft - August 2017)

Revisions to Sections on AERMOD evaluation and uncertainty, to reflect findings of November 2017 study by Hanna and Chang in which our model evaluation software, BOOT, was applied to AERMOD predictions and observations at nine samplers in Tulsa, Indianapolis and Fall River. From these calculated performance measures, we could determine model mean bias and scatter, which was not determined by EPA in their last REA draft. We had requested the data from EPA and the data were sent to Aaron Yeow and forwarded to us.

On page 3 of the 03-29-18 Draft CASAC REA Report, lines 29-30: delete sentence referring to further details in Drs. Boylan and Hanna's individual comments and replace with:

> "Following the CASAC meeting, CASAC members acquired from EPA the hourly AERMOD predictions and observations for samplers in Fall River, Indianapolis, and Tulsa, and carried out a performance evaluation with the widely-used BOOT model evaluation software. The performance was found to be typical of AERMOD performance found at many other sites and reported by EPA. That is, the fractional mean bias (FB) is 0.57, the normalized mean square error (NMSE) is 6.47, and the fraction within a factor of two (FAC2) is 0.15. Thus, for data paired by hour and location, the mean bias and scatter are about the same magnitude as the mean itself. The performance does not satisfy published criteria for dispersion models applied to rural or urban field studies. Trends in time of day (better performance for daytime), wind speed (better performance for higher wind speeds), and site (worse performance for sites in Indianapolis) are apparent. This type of evaluation should be included by EPA in their REA."

On page A-40 (Dr. Steven Hanna's individual comments), replace lines 4-14 with the following:

> "Under ambient monitor 5-minute averaged concentrations, I believe that the same nonrepresentative problem exists as for the surface meteorological data. We know that concentration variability is larger closer to the source. Also, there is a known mesoscale spatial and temporal time scale that is not being accounted for. That is, if the concentration is relatively large during one 5-minute period, it is likely to be relatively large during the next period. Typical observed mesoscale turbulence time scales are 5 to 10 minutes and space scales a few km.
>
> Under AERMOD algorithms, it should be recognized that EPA evaluations mainly consider only observations and predictions unpaired in time and space. This is why Q-Q plots are favored by EPA. If you pair the data in time and space, there is a lot of scatter (typically a factor of two), mostly due to wind direction differences. After the October CASAC meeting, I requested that EPA provide the detailed files of hourly observations and AERMOD predictions of SO_2 concentrations from the sampling sites in Fall River, Indianapolis, and Tulsa used in the CASAC

REA report. I applied the BOOT model evaluation software to these data paired in space and time.

The results are summarized in a spreadsheet and highlights are given here. As nearly always happens with evaluations of air pollution concentrations paired in time and space, the correlation is low and the scatter is high. Of course this is why EPA prefers Q-Q plots where the Co and Cp are ranked and there is no pairing in time.

In addition to calculating the performance measures for all sites combined, we present performance measures by individual sites for categories of time of day, site, and wind speed. Trends in time of day (better performance for daytime), wind speed (better performance for higher wind speeds), and site (worse performance for sites in Indianapolis) are apparent.

In our previous model evaluations with many models and field data, we find that the results can be strongly influenced by how low concentrations, threshholds and background are handled. This issue if very important here too, since the data have many low values and relatively few high values. Thus a typical mean is 3, standard deviation is 6, and highest is 150 to 200 ppb. There are many zeros in the observed concentration. When Co is 0 and Cp is non zero, then clearly Co is not within a factor of two of Cp. There is a need to discuss how to better handle the zeros and low values.

These results van be compared with our published model acceptance criteria for rural and urban field studies. Since the three sites studied here are hard to classify as urban or rural (they are in between), I list both below.

	This study	Rural acceptance criterion	Urban acceptance criterion
FB	0.57	<0.3	<0.6
NMSE	6.47	<3	<6
FAC2	0.15	>0.5	>0 3

Thus the current fractional mean bias (FB) = 0.57 does not satisfy the rural criterion and barely satisfies the urban criterion.
The current normalized mean square error (NMSE) = 6.47 does not satisfy either rural or urban criteria but barely exceeds the urban criterion of NMSE < 6
The current fraction of predictions that are within a factor of two of observations (FAC2) = 0.15 is far from satisfying the criteria. This may be because of all the Co = 0 values.

BOOT also calculated the MG and VG performance measures, which are based on lnC. This reduces the influence of a few large outliers. However because here we have some Co = 0.0 values, ln Co = - infinity! This why in our other studies we often define a threshhold C.

The main results of this evaluation of the sampler data from Fall River, Indianapolis, and Tulsa are:

1) Since FB = 0.57, it can be stated that AERMOD mean predictions are biased low by a little less than a factor of two.
2) Since NMSE = 6.47, it can be stated that the scatter is about 2.5 times the mean concentration (this is probably caused by a few large outliers)
3) Since FAC2 = 0.15, it can be stated that only 15 % of the predicted concentrations are within a factor of two of the observations.

It can be concluded that the AERMOD performance at these cities and samplers does not satisfy published criteria for dispersion models applied to rural or urban field studies.

However, this type of statistical performance is typical for AERMOD as summarized in EPA's own comparisons using many field data sets."