Atmos. Chem. Phys. Discuss., 10, C5263–C5277, 2010
www.atmos-chem-phys-discuss.net/10/C5263/2010/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Structure-activity relationships to estimate the effective Henry's law coefficients of organics of atmospheric interest" *by* T. Raventos-Duran et al.

T. Raventos-Duran et al.

aumont@lisa.univ-paris12.fr

Received and published: 13 July 2010

We wish to thank the referees for their constructive and detailed comments. We will improve our paper accordingly by addressing each of the respective points here.

Referee 1

1- As was mentioned in section 3.2, it is generally understood that multilinear regression models are prone to overfitting, hence there is a requirement for training sets in order to assess the predictive qualities of the model. There is also a necessity

to reduce the number of variables used as model parameters for the same reason. Despite this, GROMHE includes unique descriptors for many different types of substitutions, such as peracid, which has only one entry in the entire database. My question is: have the authors considered a more generalized approach to the description of substitutions? The peracid group, for example, contains a hydroperoxide group and a carbonyl group, both of which are accounted for in the database. Another example would be ketones and aldehydes, where both contain a carbonyl oxygen atom and the only difference is an alkyl substitution. Therefore, would it be possible to describe many of these substitutions using smaller fragments, which are more general and better represented in the database?

Reducing the number of descriptors as much as possible was a key aim in the development of GROMHE. Thus many attempts were made to decrease the number of descriptors. They resulted in a significant increase of the errors in the estimated values. If we describe the peracid group as a combination of ROOH and RCOR descriptors, its Henry's law constants is over-predicted by 1 order of magnitude. Similarly, the contribution obtained for the ketone group (3.29 ± 0.12) is significantly different from the one obtained for the aldehyde group (2.63 ± 0.12), where the uncertainties are the standard errors (see answer to referee 2, item 12.1 for its significance). Aggregating these two descriptors therefore leads to a significant increase of the errors. As already stated in the text (P 4621, L15), we cannot provide an assessment for the Henry's law constant estimates using GROMHE for peracid and hydroperoxide owing to the small number of entries of these type of species in the database. Nevertheless, peracid and hydroperoxyde are major functional groups produced during the tropospheric oxidation under low NOx conditions and we therefore kept these groups within GROMHE for later applications. The contribution of these groups should obviously be reviewed when additional data becomes available.

2- If possible, the authors should provide more interpretation of their multilinear regressions. There is almost no explanation of the SAR regarding the mechanism by which Henry's law constants are affected by substitution. Exactly how does the presence of an electronegative substitution neighbouring another substitution affect the Henry's law constant and why? Questions such as this can become difficult to answer as the multilinear regression becomes more complex and the physical interplay between the descriptors becomes less obvious. So, is this approach just a black box, or can some physical interpretation be made about the results of this study?

The multiple linear regression performed in this study is a fully empirical approach to estimate the Henry's law constants. Descriptors were selected when they were statistically significant for the prediction of $\log H^*$. This selection of descriptors is not based on fundamental physical properties. As stated by the reviewer, the physical interpretation of the contribution obtained for the various descriptors is not obvious, and we can only provide a very crude analysis. Dipoles and hydrogen bonds are evidently affecting the H^* of the compound but we feel that interpreting the mechanisms of such interactions is beyond the scope of this work.

3- In the final paragraph of this article a caveat is presented to the effect that the database that is used represents a subset of the available data. GROMHE was optimized for this database, but is compared with HWINb and SPARC, which are optimized for other databases. This is an inherently unfair comparison, and the general increase in performance associated with GROMHE is called into question. Since these models are complex and there may be a lot of work associated with training them, perhaps it is unreasonable to ask for these models to be re-optimized using the present database. However, is it possible to optimize GROMHE to the databases associated with HWINb and SPARC? This would result in a fair comparison and would add much more weight to any assertion that GROMHE possesses better predictive power.

As we pointed out (P4632, L4), we agree that the inter-comparison performed within this study is inherently unfair. GROMHE was developed for atmospheric applications and does not account for many functional groups found in the databases associated with HENRYWIN and SPARC methods. Adding these functional groups into GROMHE is beyond the scope of this model.

Nevertheless, the bond contributions in HENRYWIN can be reevaluated using the database described in this paper. A multiple linear regression was therefore performed to optimize the HENRYWIN model to our database. 47 descriptors (35 bonds and/or fragments and 12 correction factors) are required to describe the structure of the 488 molecules included in the database. The determination coefficient R^2 is 0.96 compared to 0.91 for the original model (a t-test shows that this difference is significant at 95 % confidence level). This optimized HENRYWIN model shows an improvement for the estimation of $\log H^*$ for the more soluble species ($H^* > 10^3$ M atm^{-1}) with an RMSE of 0.66 compared to 1.12 previously. However, GROMHE still provides a better performance. These results give confidence that our model can be used for the determination of H^* for organics known to be important in atmospheric chemistry. This comment and the figures showing the scatter plot and box plot for the optimized HENRYWIN model have been added into the revised paper.

The SPARC model uses physical parameters as descriptors (e.g. volume, molecular polarisability, molecular dipole, H bonding parameters, dispersion interaction, induction interaction, dipole-dipole interaction, H bond interaction, and entropic term). Quantum mechanics calculations are required to determine these parameters. Dealing with such tools is out of the scope of this paper and we did here not attempt to optimize SPARC for our database.

Technical corrections:
All technical corrections suggested by the referee have been implemented in the revised paper.

Referee 2

1. Most secondary organics are expected to be water soluble" (P4618 L 25): Although it is a fringe-note, this statement is ambiguous and can be misleading. Almost anything can be dissolved to some extent in water, depending on the relative amounts of water and organic present, as well as other specific (intensive) properties of the system. I suggest to give explicit (for example, order of magnitude of some property) gauges for the term "water soluble", or reformulate the statement.

We agree that this sentence can be misleading and was removed. The typical liquid water content of a cloud is about 0.3 g m^{-3} air. At equilibrium, the partitioning of a species in the aqueous phase is about 1 % if the Henry's law constant is of the order of 10^3 M atm^{-1}. In page 4620, line 19, we noted that the species of interest are those with H^* above 10^3 M atm^{-1}. Data in Table S1 show that most oxygenated multifunctional species are above that threshold.

2. I suggest chemical and mathematical equations defining central concepts are emphasized, and specifically not stated as regular text. In particular, there are different forms of Henry's law given in the literature. The version used in the present work for defining the Henry's law constant (H) is given in P4619 L 11, but should be easier to locate. This would from the beginning avoid any ambiguity concerning the definition of this central concept and immediately clarify exactly what quantity the developed model

predicts. It is especially important since the authors use the inverse form of Henry's law compared to several of the references provided (e.g. Hilal et al. (2008)).

The equation has been separated from the text.

3. The authors refer to both the Henry's law "constant" and the Henry's law "coefficient" (e.g. P4624). The view that may emerge is a distinction between the "intrinsic constant" and the "effective coefficient". If such a distinction is intended, it may be too subtle, and I suggest the different quantities are explicitly defined. If the different terminology is used for the same parameter, I suggest this ambiguity between a constant and a coefficient (that can vary as a function of various system parameters) is avoided.

There was no distinction between constant and coefficient. To avoid confusion, we changed "Henry's law coefficient" into "Henry's law constant" in the text (including the title).

4. As the authors note, the version of Henry's law used for defining the Henry's law constant is a limiting law (P4619 L 13), without activity coefficient correction and thus pertaining to the infinite dilution state of the solute. An underlying assumption of the SPARC method (Hilal et al., 2008) is that it also applies in the solubility limit of a saturated solution (P4620 L 10). As mentioned in several of the references provided, the variability or uncertainty related to Henry's law constants measured by composite methods is in some cases related to confusion of states and the extrapolation of Henry's limiting law to the solute solubility limit. A clarifying comment about the uniformity of the reference state assumptions involved at the different stages in the model development and the significance of the predictions with GROMHE to atmospheric *applications (where aqueous aerosol solutions are rarely dilute) might be useful. In fact, the discussion might benefit from adding a short theory section clarifying these ambiguities.*

We have the feeling that adding a "theory section" is beyond the goal of this paper. We added a reference to Boethling and Mackay, 2000, where further details can be found. Reference: Handbook of property estimation methods for chemicals: environmental and health sciences, R.S. Boethling and D. Mackay, Lewis Publishers, CRC Press, New York, 2000.

5. The authors might consider spending a few lines describing the difference between the GROMHE, HWINb and SPARC models, since the performance of GROMHE in comparison to these existing tools is a main result presented. What the significant new contribution of the GROMHE, compared to the HWINb and SPARC? Besides the argued improvement of statistics, any new concepts should be more explicitly emphasized to clarify this further.

The following comment has been inserted in the introduction (P4620, L23):
The SARs studied in this work are all based on a multiple linear regression approach. The main difference relies on the selection of descriptors (i.e. the predictors) used to estimate H^*. The descriptors chosen by SPARC are physical parameters (e.g. volume, molecular polarisability, molecular dipole, H bonding parameters, dispersion interaction, induction interaction, dipole-dipole interaction, H bond interaction, entropic term, etc) and quantum mechanic calculations are required to determine their values (Hilal et al., 2004). HWINb uses simple molecular structural descriptors: the number and type of the chemical bonds and in addition, some correcting factors. GROMHE uses a similar paradigm as HWINb, but is based on the number and nature of the functional groups present in the molecule.

6. P4621: The authors point out that the data set is very limited for some compound classes and this affects the reliability and/or predictive ability of the model for other such compounds. I understand that the purpose of this work is to develop the GROMHE tool from already existing data. Nevertheless, it would be most useful for targeting future experimental work if the authors could provide recommendations and/or guidelines for measurements that would specifically facilitate further optimization of the model developed.

See answer to referee 1, item 1. In addition, the following sentences have been added in the paper (P4621, L14): The availability of experimental data for hydroperoxides (3 species) and peracids (1 species) is limited and therefore it is difficult to assess the reliability of H^* estimates for these groups of species. This is a limiting factor since oxidation proceeds trough the formation of such compounds in remote conditions (or low NOx conditions). Additional data for these groups of species would be especially valuable to constrain structure activity relationships for atmospheric applications.

7. P4622, L7: "The data were exclusively taken from experimental values either from direct or indirect measurements." As opposed to what - which other forms of experimental values could have been used? I have an idea what the authors mean, but perhaps this could be clarified.

The word "exclusively" is misleading and has been removed from the sentence.
This sentence refers to 'experimental measurements' used as opposed to data from "modelling studies" which can be found listed in database compilations (e.g. see reference in the paper Sander, 1999).

8. P4622, L21: What is the sensitivity of the model to the estimated variation in the desolvation enthalpy (ΔH$_{\text{solv}}$)? Although I understand that the experimental uncertainty for H is generally large, it would be good if the authors could provide a remark about whether such sensitivity is comparable to or minor in comparison with these uncertainties, and for how many compounds the temperature extrapolation is relevant.

As described in page 4622, line 15, the experimental data used to develop the model are provided at 298 K. A small number of species (20 compounds) measured at 293 K were also included to obtain a better representation of multifunctional oxygenated species. Values at 293 K were corrected using the Van't Hoff equation. The desolvation enthalpy ΔH$_{\text{solv}}$ must be known to apply this correction. Here, we assume a typical value of 50 kJ mol^{-1} for all species. For a 5 K change in temperature (293 → 298 K), this value of ΔH$_{\text{solv}}$ leads to a 30 % decrease of H^*. For species measured at 293 K, the H^* values in Table S1 was thus decreased by 0.15 log units. The desolvation enthalpy ΔH$_{\text{solv}}$ typically ranges from 10 to 100 kJ mol^{-1} (e.g. Kuhne et al., 2005). This span of enthalpies leads to a decrease of H^* ranging from 7 to 50 % for a 5 K increase (i.e. from 0.03 to 0.3 log unit). The uncertainty in the correction factor for the 293 → 298 K change is small compared to the uncertainties in the model outputs and experimental data.

9. The model predicts H at 298 K, but how sensitive are model predictions with GROMHE to atmospherically relevant temperatures? What do the authors believe is the valid temperature range of the model? What is considered to be the relevant temperature range for atmospheric applications? For example, how would the model perform for, say, 278 K?

As discussed above (item 8), the value of H^* are temperature sensitive. Assuming

a value of 50 kJ mol^{-1} for the desolvation enthalpy, a temperature change from 298 to 263 K leads to a decrease of about 1 order of magnitude on the value of H^*. To predict Henry's law constants at other temperatures, the model developed here must be coupled to an additional model that estimates the values of desolvation enthalpies (e.g. see reference in the paper Kuhne et al., 2005).

10. P4624: The authors address hydration of carbonyls and the effect on observed effective Henry's law coefficients, as well as the means to account for this effect in the developed model. What about the effects of other solute-solvent equilibria, e.g. hydrolysis reactions of acid functionalities and protonation of bases?

We took into account the hydration processes for carbonyls because it is usually implicitly accounted in the measured Henry's law constant. Acid/base equilibrium constants are another type of key properties that also must be estimated for atmospheric applications. SARs to estimate pKa for organic acids and bases are already existing in the literature (e.g. see reference in the paper Perrin et al.,1981).

11. P4624 L18: This paragraph starts very abruptly and it is difficult to follow. The purpose of the paragraph is not immediately clear, if you are not already familiar with the concepts, as it introduces and discusses new quantities without much explanation. For example, what is a "descriptor", and what is the purpose of introducing it? I was also confused about the sudden discussion of aromatics, aliphatics, then aldehydes and ketones: Maybe "Taft and Hammett_" values (P4624 L22) could be briefly defined and put into context?

An introduction to the paragraph and further explanations have been added:
A SAR was constructed to estimate K_{hyd} based on a multiple linear regression using

the experimental data shown in Table S2 as training set. This modeling approach assumes that the relationship between the dependent variable y (here K_{hyd}) and the independent variables x_j (here the structural descriptors or predictors) is linear. The equation for this model is given by:

$$y_i = \beta_0 + \beta_1 x_{1,i} + ... + \beta_j x_{j,i} + ... + \beta_p x_{p,i} \tag{1}$$

where i stands for the i^{th} species in the database, $x_{j,i}$ is the j^{th} descriptors and β_j are the regression coefficients (here the weights or contributions) computed for the descriptor j. The best-fitting line for the observed data is calculated by minimizing the sum of the squared errors, SSE:

$$\text{SSE} = \sum_{i=1}^{n} (\log K_{hyd,est} - \log K_{hyd,exp})^2 \tag{2}$$

where n is the number of species included in the database. The descriptors were selected following their assessment in the multiple linear regression.

12. The greatest difficulty for me in assessing the methods and results presented by the authors concerns the statistical methods and parameters for which values are provided. Although it may be argued that this should ideally be basic knowledge to everyone working in science, at least I lack the proper knowledge foundation to evaluate the quality and performance of the developed model. Specifically,

12.1. What is precisely meant by "reliability", "quality", "prediction ability", "contribution", and "weight in the regression"? What are the definition and units of the "standard errors" of the descriptors?
- *Reliability*: in the text, reliability was used to express the magnitude of the error associated to the estimations. A more reliable method or estimation has lower errors.

- *Quality and prediction ability*: the sentence using it (P4625, L22) was removed to avoid confusion and rephrased as follow using referee 1 (item 1) suggestion. The identification of descriptors for the multiple linear regression is complex: increasing the number of descriptors (increase in the degree of freedom) usually leads to a better fit of the experimental data. However, regression models are prone to over-fitting and there is a requirement to reduce the number of descriptors used as much as possible.
- *Contribution and weight in regression*: they have now been defined in P4620, (see item 11). "Weight in the regression" was deleted when rephrasing the sentences. "Weight" was changed to contribution (or regression coefficient).
- *Definition and units of the "standard errors" of the descriptors*: The standard errors (in Tables 1 and 3) have the units of the descriptor's contribution. It represents the standard deviation of the j^{th} regression coefficient (or contribution). It can be used to calculate the confidence interval of each contribution as:

$$\beta_j = \hat{\beta}_j \pm t_\alpha \times se_j \tag{3}$$

where $\hat{\beta}_j$ is the estimated value of the j^{th} contribution (given in Tables 1 and 3), se_j is the standard error, and t_α is the student number for a given threshold or level of significance α and for the degree of freedom (df) in our regression analysis:

$$df = n - k - 1 \tag{4}$$

where n is the total number of species (here 488), k is the number of descriptors (here 28). For example, the contribution calculated for the number of carbon atoms is 0.49 with a se of 0.02 (see Table 1). For $\alpha = 0.025$, then $t = 1.96$. Therefore, at 95 % confidence, we can say that the contribution of that descriptor is between 0.45 and 0.53.

12. 2. What is the significance/interpretation of the different types of errors discussed, i.e. the RMSE, MAE, MBE (P4625, P4628)? How do these quantities assess the *model and method? What does a certain value mean? What is "good" or "bad"? Also, the "n" in the defining equations is unspecified?*

We think that the formulas provided to define the errors are self-explanatory. In short, the MBE allows to check whether the predictions show some bias (i.e. a systematic overestimation or underestimation for the whole set or a subset of the database). The RMSE can be viewed as the standard deviation of the error (i.e. if the distribution of the errors follow a normal distribution, 68 % of the predicted values are within plus or minus the RMSE). For example, we show that GROMHE estimates $\log H^*$ with a RMSE of 0.53 for species having H^* above 10^3 M atm^{-1}. If we assume that the distribution of the errors are close to a normal distribution (a reasonable approximation given the shape of the box plot shown in Figure 4), this means that for this subset, about 68 % of the estimated values are within 0.53 log units of the experimental ones and 95 % within 1.06 log units (2×RMSE). The MAE is typically used to measure the average magnitude of the errors in a set of predictions. The MAE is a linear index which means that all the individual errors are weighted equally in the average. In contrast, the RMSE is a quadratic index which gives relatively high weight to large errors.

We introduced 'n' (the number of experimental data) in page P4620 of the text and change 'N' for 'n' in the graphs.

Good or bad: An error that falls within the uncertainty of the experimental data would appear as "very good". As stated in the text, the uncertainty in the experimental data is at least a factor of 2 (page 4623, line 6) for species with large H^* values. Therefore, we think that the estimated values can be stated as "good" if they agree within a factor of 4 (or 0.6 log units) of the experimental one. The estimations using GROMHE show errors of the same order of magnitude.

12. 3. It seems like a "Catch 22" that the GROMHE is evaluated in comparison with other models against the data set on the basis of which it was developed? I see this is noted by the authors themselves in the conclusions section (P4632 L5), but again, some remark elaborating a bit further on the significance of this could be very informative.

See answer to referee 1, comment 3.

13. I believe it would be a great benefit if the central statistical parameters and their interpretation was explained shortly, since it is the foundation of the development and evaluation of the model described. It might furthermore enable the reader to better appreciate the significance and quality of the results. Naturally, the authors cannot write a new statistics text book, but a little more basic information might greatly improve readability to people outside the specific field, notably the end users of the model developed. This would enable the users to evaluate the strengths and weaknesses of the model for the various application purposes. Such a statistics summary could be included as supporting material or an appendix.

The statistical concepts used in this paper can be found in most statistical textbooks. We think it will be more useful for the reader to refer to this specialized literature for more details.

Technical Comments
Tables in supporting material (1, 2): Regarding the typography of the molecular formulae, the authors could 1. use the "back-slash{chem}" command supported in the Copernicus LaTeX package. 2. capitalize the appropriate elemental designation

letters. 3. clarify what is meant by "SMILES" (table T1 and T2, leftmost columns' headings): at least I am not familiar with this term.

The molecular formulas in table S2 are written in the SMILES formatting system. This system is used in many chemical computer softwares to provide the formula as a character strings. More details on this can be found in http://www.daylight.com/dayhtml_tutorials/languages/smiles/index.html Table S2 provides the CAS numbers to avoid confusion.

Other technical corrections suggested by the referee have been performed in the revised paper.

Interactive comment on Atmos. Chem. Phys. Discuss., 10, 4617, 2010.