Structure-activity relationships to estimate the effective Henry's law constants of organics of atmospheric interest

T. Raventos-Duran, M. Camredon, R. Valorso, C. Mouchel-Vallon, B. Aumont

Atmos. Chem. Phys., 10, 7643–7654, 2010
www.atmos-chem-phys.net/10/7643/2010/
doi:10.5194/acp-10-7643-2010

Atmospheric Chemistry and Physics

Structure-activity relationships to estimate the effective Henry's law constants of organics of atmospheric interest

T. Raventos-Duran, M. Camredon, R. Valorso, C. Mouchel-Vallon, and B. Aumont

LISA, UMR CNRS/INSU 7583, Université Paris Est Créteil et Université Paris Diderot, Institut Pierre Simon Laplace, 94010 Créteil Cedex, France

Received: 19 January 2010 – Published in Atmos. Chem. Phys. Discuss.: 16 February 2010
Revised: 16 July 2010 – Accepted: 26 July 2010 – Published: 17 August 2010

Abstract. The Henry's law constant is a key property needed to address the multiphase behaviour of organics in the atmosphere. Methods that can reliably predict the values for the vast number of organic compounds of atmospheric interest are therefore required. The effective Henry's law constant H^* in air-water systems at 298 K was compiled from literature for 488 organic compounds bearing functional groups of atmospheric relevance. This data set was used to assess the reliability of the HENRYWIN bond contribution method and the SPARC approach for the determination of H^*. Moreover, this data set was used to develop GROMHE, a new Structure Activity Relationship (SAR) based on a group contribution approach. These methods estimate $\log H^*$ with a Root Mean Square Error (RMSE) of 0.38, 0.61, and 0.73 log units for GROMHE, SPARC and HENRYWIN respectively. The results show that for all these methods the reliability of the estimates decreases with increasing solubility. The main differences among these methods lie in H^* prediction for compounds with H^* greater than $10^3\,\mathrm{Matm}^{-1}$. For these compounds, the predicted values of $\log H^*$ using GROMHE are more accurate (RMSE = 0.53) than the estimates from SPARC or HENRYWIN.

1 Introduction

The oxidation of hydrocarbons emitted in the atmosphere involves complex reaction sequences. This oxidation is a gradual process leading to the formation of oxygenated organic intermediates usually denoted as secondary organics (e.g., Atkinson, 2000). The fate of these secondary organics remains poorly quantified due to a lack of information about their speciation, distribution and evolution in the gas and condensed phases (e.g., Goldstein and Galbally, 2007). A significant fraction of secondary organics may dissolve into the tropospheric aqueous phase, namely rain, clouds and deliquescent particles (e.g., Saxena and Hildemann, 1996; Facchini et al., 1999). The resulting mass transfer is currently suggested to contribute to acid production, organic aerosol formation and the oxidant budget (e.g., Lelieveld and Crutzen, 1990; Walcek et al., 1997; Blando and Turpin, 2000; Ervens et al., 2003, 2008; Legrand et al., 2003, 2005; Yu et al., 2005; Gelencser and Varga, 2005; Lim et al., 2005; Hallquist et al., 2009).

In atmospheric models, the partitioning of organics between the gas and the aqueous atmospheric phases is usually described in the basis of Henry's law (e.g., Jacob et al., 1989; Aumont et al., 2000; Herrmann et al., 2000, 2005; Ervens et al., 2003, 2008; Pun et al., 2002; Griffin et al., 2003). Henry's law expresses the relationship between the solubility of a gas in a liquid and its partial pressure above that liquid:

$$S = H \times P \tag{1}$$

where S is the solubility (M), P is the partial pressure (atm) and H is the Henry's law constant (Matm^{-1}) at a given temperature. Henry's law is a limiting law that strictly applies to ideally dilute solutions (e.g., Levine, 2002; Boethling and Mackay, 2000). Atmospheric models require a knowledge of H for every water soluble organic species described in the chemical mechanism. Detailed gas phase or multiphase chemical mechanisms involve a vast number of species (e.g., Saunders et al., 2003; Aumont et al., 2005; Herrmann et al., 2005). The collection of Henry's law constants required to develop detailed models far exceeds the number of species for which experimental data is available. For example, the fully explicit oxidation mechanism developed by Camredon et al. (2007) for 1-octene includes 1.4×10^6 species and the gas/aerosol thermodynamic equilibrium for about 4×10^5 species. Reliable estimation methods for H are therefore required to design detailed mechanisms. To

Correspondence to: B. Aumont
(aumont@lisa.u-pec.fr)

be useful, estimation methods must be applicable to a wide range of organics, especially to multifunctional species generated during the atmospheric oxidation of hydrocarbons. The aim of this paper is to identify a reliable method for estimating Henry's law constants for organic compounds of atmospheric interest in air-water systems.

Numerous structure activity relationships (SARs) have been developed to determine the Henry's law constants in a response to the difficulties associated with its laboratory measurement, in particular, for compounds with higher solubility (Mackay and Shiu, 1981; Russell et al., 1992; Hine and Mookerjee, 1975; Meylan and Howard, 1991; Suzuki et al., 1992). These SARs were reviewed and analysed by Dearden and Schuurmann (2003). This study showed that the bond contribution method developed by Meylan and Howard (1991) and updated in the frame of the HENRYWIN (HWINb) software (Meylan and Howard, 2000) was the most reliable method available. Dearden and Schuurmann (2003) analysed 700 compounds with HWINb and found a R^2 of 0.88 with a standard deviation of 1.03. Recently, a new method, SPARC, has been developed by Hilal et al. (2008). This method is based on the product of the activity coefficient in water γ_w^∞ and the vapour pressure P^o which are estimated using intermolecular interactions in the pure liquid phase and in solution (e.g., Boethling and Mackay, 2000). Hilal et al. (2008) used an experimental database of 1222 compounds to test the air-to-water H. Their results show that for simple molecular structures, the standard deviation is within a factor of 2 but reaches a factor of 3 to 4 for more complex molecules having strong intramolecular and/or dipole-dipole interactions.

The objective of this paper is to assess the reliability of HWINb and SPARC methods. To this end an experimental database of Henry's law constants was compiled. Special attention was given to select those compounds with H above $10^3\,\mathrm{Matm}^{-1}$ which are soluble enough to have significant partitioning in the atmospheric aqueous phase (e.g., Seinfeld and Pandis, 1997; Gelencser and Varga, 2005). Furthermore, this database was used to develop a new SAR: the GROup contribution Method for Henry's law Estimate (hereafter named GROMHE). HWINb, SPARC and GROMHE are all SARs based on a multiple linear regression approach. The main difference between them 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, H bond interaction, entropic term, etc.) and quantum mechanical 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 follows a similar paradigm but is based on the number and nature of the functional groups present in the molecule (see Sect. 3.2).

In this paper, we first describe the selection of the database used to develop and/or assess the estimation methods. We then describe the development of GROMHE and finally analyse the performance of the three methods considered for this study.

2 Database

Usually the experimental values found in the literature are expressed as effective Henry's law constants, H^*, which includes the hydration process. We differentiate the literature H^* values from the intrinsic H values as detailed in the next section. The database of Henry's law constants was compiled to include species representative of atmospheric oxidation processes occurring in the gas or aqueous phase. Table S1 (see the electronic supplement) lists the experimental values selected in this study in units of Matm^{-1} and presented as the logarithm of H^*. The database includes 488 organic compounds comprising a wide range of functional groups detected in either gas or aqueous phase: nitrate, nitro, peroxyacylnitrate, aldehyde, ketone, ester, ether, alcohol, hydroperoxide, peracid, carboxylic acid and halogen (e.g., Finlayson-Pitts and Pitts, 2000; Seinfeld and Pandis, 1997). The number of species bearing a specific functional group is given in Table 1. The availability of 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 through the formation of such compounds in remote conditions (or low NO_x conditions). Additional data for these groups of species would be especially valuable to constrain structure activity relationships for atmospheric applications. The database is also poor for multifunctional oxygenated organics, although special care was taken to be as comprehensive as possible in the collection of experimental H^* for these groups of species. Data listed in Table S1 includes H^* for 76 hydrocarbons, 231 monofunctional compounds, 132 difunctional compounds and 49 compounds bearing at least 3 functional groups. Both aliphatic and aromatic species were considered in the compilation and the data in Table S1 can be split into 393 aliphatic and 95 aromatic species. The constants included range from 10^{-4} to $10^9\,\mathrm{Matm}^{-1}$. Henry's law constants depend on the type of functional groups attached to the carbon chain and usually increase with the number of groups; for hydrocarbon species H^* ranges from 10^{-4} to $10^{-1}\,\mathrm{Matm}^{-1}$ whilst for monofunctional organic compounds H^* ranges from 10^{-1} to $10^5\,\mathrm{Matm}^{-1}$. Difunctionals compounds have the greatest range of H^*, from 10^{-1} to $10^9\,\mathrm{Matm}^{-1}$.

Most of the Henry's law constants used in this study were collected from three different libraries; NIST (http://webbook.nist.gov/chemistry), the Sander data review (http://www.mpch-mainz.mpg.de/sander/res/henry.html), and the

Table 1. Descriptors for the model GROMHE, number of species in the database contributing to the descriptor and their related contribution, standard error and statistical significance in the MLR.

Descriptor[a]	Training dataset				All dataset			
	Number of species	Contribution	Standard Error	*p*-Value	Number of species	Contribution	Standard Error	*p*-Value
Functional group and structural descriptors								
# of hydroxy groups (-OH)	85	4.56	0.11	0.0000	120	4.56	0.09	0.0000
# of nitro groups (-NO2)	22	3.06	0.12	0.0000	27	3.02	0.10	0.0000
# of nitrate groups (-ONO2)	33	2.06	0.07	0.0000	44	2.04	0.06	0.0000
# of hydroperoxide groups (-OOH)	1	4.98	0.42	0.0000	3	4.87	0.24	0.0000
# of fluorine groups (-F)	15	0.60	0.10	0.0000	19	0.60	0.08	0.0000
# of chlorine groups (-Cl)	26	0.88	0.07	0.0000	51	0.87	0.06	0.0000
# of bromine groups (-Br)	15	1.04	0.10	0.0000	21	1.06	0.09	0.0000
# of iodine groups (-I)	7	1.15	0.18	0.0000	11	1.22	0.13	0.0000
# of aldehyde groups (-CHO)	18	2.63	0.12	0.0000	24	2.59	0.11	0.0000
# of ketone groups (-COR)	22	3.29	0.12	0.0000	35	3.16	0.10	0.0000
# of acid groups (-COOH)	27	5.11	0.11	0.0000	36	5.09	0.09	0.0000
# of peracid groups (-COOOH)	1	4.68	0.41	0.0000	1	4.68	0.40	0.0000
# of peroxyacyl nitrate groups (-PAN)	3	1.94	0.25	0.0000	5	1.93	0.19	0.0000
# of ether groups (-OR)	42	2.44	0.10	0.0000	52	2.40	0.09	0.0000
# of ester groups (-COOR)	37	2.79	0.10	0.0000	55	2.78	0.08	0.0000
# of formate groups (-HCOOR)	3	2.39	0.25	0.0000	4	2.36	0.21	0.0000
# of C atoms	345	0.49	0.02	0.0000	488	0.50	0.02	0.0000
# of H atoms	345	−0.31	0.01	0.0000	488	−0.31	0.01	0.0000
nfcd	26	−0.59	0.07	0.0000	37	−0.52	0.05	0.0000
nfaro	48	−1.10	0.07	0.0000	67	−1.12	0.06	0.0000
Group interaction descriptors								
tdescriptor	98	−0.14	0.01	0.0000	138	−0.14	0.01	0.0000
caox-a	9	−1.78	0.17	0.0000	13	−1.77	0.13	0.0000
caox-b	8	−1.31	0.18	0.0000	12	−1.09	0.14	0.0000
hyd-a	18	−0.63	0.13	0.0000	29	−0.60	0.10	0.0000
hyd-b	15	−1.00	0.18	0.0000	23	−1.03	0.14	0.0000
Correction factor descriptors								
haloic-a	5	0.98	0.21	0.0000	10	0.97	0.15	0.0000
onitrofol	7	−2.72	0.23	0.0000	10	−2.66	0.19	0.0000
nogrp	52	−0.31	0.11	0.0069	76	−0.28	0.09	0.0028
Intercept	–	−1.51	0.11	0.0000	–	−1.52	0.09	0.0000

[a] See Sect. 3.2 for the meaning of the descriptor.

Environment Protection Agency HENRYWIN program (Meylan and Howard, 2000) with a few additional values taken from recently published papers (see Table S1 in the electronic supplement). The data were taken from experimental values either from direct or indirect measurements. The indirect measurements are based on relationships between thermodynamic variables. In particular, for sparingly water soluble species, H^* is often estimated using the relationship: $H^* = S_w^s / P^o$ where S_w^s is the solubility for a saturated solution and P^o is the vapour above the pure compound in the condensed phase. Because S_w^s and P^o values are measured independently in the laboratory, we have two sources that contribute to the uncertainty in the final H^* value (Mackay and Shiu, 1981).

Most experimental H^* data in Table S1 are provided at 298 K. A small number of species measured at 293 K (20 compounds, see Table S1) were also included to obtain a better representation of multifunctional oxygenated species. These values were corrected using the van't Hoff equation:

$$H_{298} = H_{293} \times \exp\left(\frac{\Delta \mathrm{H_{solv}}}{R}\left(\frac{1}{293} - \frac{1}{298}\right)\right) \quad (2)$$

where $\Delta \mathrm{H_{solv}}$ is the desolvation enthalpy and R is the gas constant. The desolvation enthalpy $\Delta \mathrm{H_{solv}}$ 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 units). Here, we assume a typical value of 50 kJ mol^{-1} for all species. For each species measured at 293 K, the value of H^* in Table S1

was thus decreased by 0.15 log units. The uncertainty in the applied correction factor is small compared to the uncertainties from the SAR outputs and experimental data.

For empirically based methods, the experimental uncertainties are transferred into the models' uncertainties. However, uncertainties for the data reported in Table S1 are hard to evaluate owing to the large number of experimental sources and the lack of reported experimental uncertainties in many of the original papers (Russell et al., 1992). During the compilation, we found that discrepancies in the measured H^* value for a given compound often exceeded a factor of 2. The discrepancies tend to increase with H^* which indicates the difficulties of measuring the physical property for those species with very high Henry's law constants (Hilal et al., 2008; Mackay and Shiu, 1981). As a guideline, we assume that the uncertainty is at least a factor of 2 for species having H^* above $10^5\,\mathrm{Matm}^{-1}$.

3 Development of the GROMHE estimation method

3.1 Estimation method for hydration constants

Compounds containing carbonyl groups like aldehydes and ketones may undergo significant hydration. Carbonyls combine with water molecules to form gem diols upon dissolution according to the equilibrium:

$$> C = O + H_2O \quad \longleftrightarrow \quad > C(OH)(OH)$$

The equilibrium of the carbonyls between the hydrated ($> C(OH)(OH)$) and non-hydrated ($> C{=}O$) form is described by the hydration constant K_{hyd}:

$$K_{\mathrm{hyd}} = \frac{[> C(OH)(OH)]}{[> C{=}O]} \tag{3}$$

Table S2 (see the electronic supplement) shows the compilation of the hydration constants for 61 aldehydes and/or ketones. K_{hyd} is typically about 10^{-3} and 1 for simple ketones and aldehydes, respectively. K_{hyd} increases by 1 to 3 orders of magnitude when a strongly polar group is attached to the carbon atom next to the carbonyl group. Hydration is therefore a key parameter affecting the solubility of multifunctional carbonyl compounds.

The partitioning of species that undergo hydration in water is usually described with the effective Henry's law constant H^*. The effective Henry's law constant of a compound is defined as the ratio between the total dissolved concentration and its pressure:

$$H^* = \frac{([> C{=}O] + [> C(OH)(OH)])}{P_{> C{=}O}} = H\left(1 + K_{\mathrm{hyd}}\right) \tag{4}$$

where H is the intrinsic Henry's law constant for the carbonyl. The values extracted from the literature and listed in Table S1 (see the electronic supplement) are therefore H^* for carbonyls.

Most estimation methods were based on group contribution methods using $\log H^*$ as the training data set. Equation (4) shows that for carbonyls, H^* is a function of 2 fundamental properties (H and K_{hyd}). If $K_{\mathrm{hyd}} \gg 1$ then $\log H^* \approx \log H + \log K_{\mathrm{hyd}}$. On the other hand, if $K_{\mathrm{hyd}} \ll 1$, then $\log H^* \approx \log H$. This conditional addition of $\log K_{\mathrm{hyd}}$ is hard to represent by a simple group contribution method which assumes additive groups. Here we estimated both K_{hyd} and the intrinsic H for each carbonyl and used Equation 4 to finally compute H^*. Note that the method performance was assessed on the accuracy of H^* which is the primary property being investigated.

A SAR was constructed to estimate K_{hyd} based on a multiple linear regression approach using the experimental data shown in Table S2 as training set. This modelling approach assumes that the relationship between the dependent variable y_i (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:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_j x_{j,i} + \cdots \tag{5}$$

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 contributions) computed for the descriptor j. The best-fitting line for the observed data is calculated by minimising the sum of the squared errors, SSE:

$$\mathrm{SSE} = \sum_{i=1}^{n} \left(\log K_{\mathrm{hyd,est}} - \log K_{\mathrm{hyd,exp}}\right)^2 \tag{6}$$

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

Previous studies have shown that K_{hyd} is well correlated with the inductive effect of the neighbouring groups (Le Henaff, 1968; Betterton and Hoffmann, 1988). Therefore, Taft and Hammett σ were used as descriptors for aliphatic and aromatic compounds, respectively (see Table 2). Hammett values for the various functional groups were obtained from the data reviews of Hansch et al. (1995) and Perrin et al. (1981). We defined a "Hammett descriptor" (referenced as *hdescriptor* in Table 3) as the sum of the contribution of each group:

$$hdescriptor = \Sigma\sigma_o + \Sigma\sigma_m + \Sigma\sigma_p \tag{7}$$

where σ_o, σ_m, σ_p are the Hammett sigma values for the functional groups in ortho, meta or para positions relative to the benzaldehyde group (see Table 2). Similarly, we defined a "Taft descriptor" (*tdescriptor*) as:

$$tdescriptor = \Sigma\sigma_i^* \tag{8}$$

where σ_i^* are the Taft sigma values for the functional groups i borne by the molecule in relation to the carbonyl group (see Table 2). Two additional molecular descriptors were introduced to discriminate aromatic from aliphatic compounds

Table 2. Sigma Taft and Hammett values for organic functional groups (adapted from Perrin et al., 1981).

Functional group	Taft σ*[a]	Hammett ortho σ_o	Hammett meta σ_m	Hammett para σ_p
ROH	0.62	0.13	−0.38	1.22
RNO_2	1.47	0.74	0.78	1.99
$RONO_2$[b]	1.54	0.55	0.7	–
ROOH[c]	0.62	–	–	–
RF	1.10	0.34	0.06	0.93
RCl	0.94	0.37	0.24	1.28
RBr	1.00	0.39	0.22	1.35
RI	1.00	0.35	0.21	1.34
RCHO	2.15	0.36	0.44	0.36
RCOR	1.81	0.36	0.47	0.07
RCOOH	2.08	0.35	0.44	0.95
COOOH[c]	2.08	–	–	–
PAN[c]	2.00	–	–	–
ROR	1.81	0.11	−0.28	0.12
ROCOR[d]	2.56	0.32	0.39	0.63
RCOOR[e]	2.00	0.32	0.39	0.63
HCOOR	2.90	–	–	–

[a] Reported σ^* is the inductive effect that the carbon bearing the functional group exerts on its direct neighbouring groups. According to Perrin et al. (1981) σ^* for functional groups attached to carbons at distant positions are determined as $\sigma = \sigma_i \times (0.4)^n$ where n is the number of aliphatic carbons separating the functional groups. The 2 carbons forming a C=C bond are counted as one C only. [b] Perrin et al. (1981) gives $\sigma^* = 3.86$ for the nitrate group. The value reported here is $\sigma^* = 0.4 \times 3.86$, estimated for the carbon bearing the nitrate functional group to its neighbouring groups. [c] Value set assuming that $\sigma_{ROOH} = \sigma_{ROH}$, $\sigma_{C(O)OOH} = \sigma_{RC(O)OH}$, $\sigma_{PAN} = \sigma_{RCOOCH3}$. [d] Sigma for ester at the -O- side. [e] Sigma for ester at the -CO side.

and ketone from aldehyde groups. Table 3 provides the optimised contribution for these 4 descriptors and Fig. 1 shows the resulting scatter plot. The coefficient of determination is $R^2 = 0.91$. The reliability of the method was assessed using the root mean square error (RMSE) defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log K_{\mathrm{hyd,est}} - \log K_{\mathrm{hyd,exp}}\right)^2} \quad (9)$$

where n is the number of species included in the database. The RMSE obtained is 0.47 log units. The method allows estimating K_{hyd} within a factor of 3.

3.2 Estimation method for the intrinsic Henry's law constants

The GROMHE approach is similar to the method described by Suzuki et al. (1992) and is based on considering a molecule as a collection of elemental constituents (functional groups or atoms) whose contributions are computed using a multiple linear regression (MLR). The original approach by Suzuki et al. (1992) was developed for monofunctional species only. In GROMHE, the approach is extended to multifunctional species using additional descriptors to account for group interactions. 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. An attempt was thus made to minimise the number of descriptors and to optimise the regression for the species of atmospheric interest.

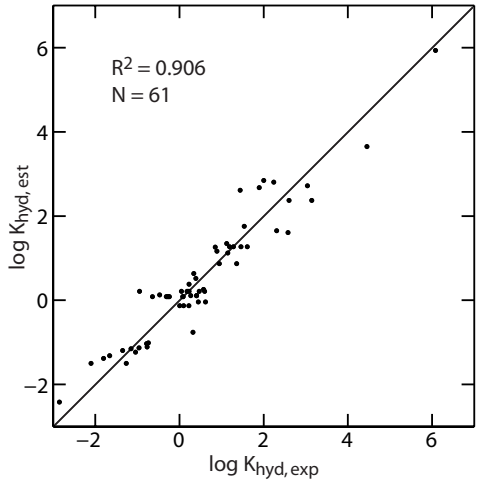

Fig. 1. Estimated hydration constant versus experimental values. The line is the $y = x$ line.

The database was split into two sets: 70% of the data were used as training set and the remaining 30% were reserved for validation and were not used during the development of the method. The training data set was then used to compute

Table 3. Descriptors for the model to estimate hydration constants and their related contribution, standard error and statistical significance (p-value) in the MLR.

Descriptor	Contribution	Standard Error	p-Value
tdescriptor [a]	1.27	0.07	0.0000
hdescriptor [a]	0.50	0.17	0.0049
Ketone flag [b]	−2.50	0.17	0.0000
Aromatic flag [b]	−1.58	0.24	0.0000
Intercept [b]	0.08	0.12	0.4968

[a] See Sect. 3.1. [b] Flag is Boolean type set to 1 if the criterion is matched.

the contribution of the descriptors selected for the regression. Species used for validation were randomly selected and are given in Table S1 (see the electronic supplement). This random selection covered structurally diverse compounds representative of all type of functional groups included in the database (see Table 1). The effective Henry's law constants collected in our data set were corrected for hydration to determine intrinsic values. The structure activity relationship (SAR) presented in the previous section was used systematically to derive the hydration constants for ketones and aldehydes and to compute the intrinsic Henry's law constant (H) values. These derived intrinsic H values were used as the training data set for the MLR analysis. Our model uses 28 independent descriptors, presented below. The list of descriptors along with their contributions and standard errors are shown in Table 1.

Suzuki et al. (1992) have shown that H can be estimated for hydrocarbons and monofunctional species using the organic functionalities as descriptors along with the number of carbon and hydrogen atoms. We introduced 16 descriptors, each corresponding to a distinct organic functionality identified within the compounds comprising the study data set (see Table 1), and two structural descriptors to account for the number of hydrogen and carbon atoms.

In contrast to Suzuki et al. (1992) who duplicated the descriptors to differentiate functionalities bound to an aromatic chain from those bound to an aliphatic chain, we simply defined two additional descriptors to account for the number of groups bound to an aromatic ring or an olefinic carbon respectively, so as to keep the number of descriptors to a minimum. These descriptors are referenced as *nfaro* and *nfcd* in Table 1. An MLR using these 20 structural descriptors was able to provide H estimates with an R^2 of 0.97 for the hydrocarbons and monofunctional species included in the data set.

Extrapolation of our model using the 20 descriptors defined above however leads to errors in the estimated values exceeding 3 orders of magnitude for some difunctional species. Additional descriptors were therefore included to account for intramolecular group interactions. The mutual inductive effect between functional groups was explored as a parameter linked to the overestimation of H identified in multifunctional species. Here, we introduced Sigma Taft (σ^*) as a descriptor for aliphatic species (e.g., Hansch et al., 1995). Group interactions were taken into account by adding, for each group i, the σ_j^* of the neighbouring group j:

$$tdescriptor = \sum_i \sum_{j \neq i} \sigma_j^* \quad (10)$$

where $tdescriptor$ is the parameter used as a descriptor for the regression. Values for σ_j^* are provided in Table 2 for each of the 16 functional groups encountered in the database. $tdescriptor$ was found to be statistically significant for the prediction of H at the 99.9% confidence level (see p-value in Table 1). The inclusion of $tdescriptor$ in the set of descriptors leads to a fairly good estimate of H for multifunctional compounds bearing nitro, nitrate and/or halogen groups. However, H was still overestimated for multifunctional species bearing carbonyl or hydroxyl moieties, so additional descriptors were introduced.

Scatter plots showed that species with a $-C(=O)-C(X)<$ structure where X is an oxygenated moiety (carbonyl, alcohol, ether, hydroperoxide or nitro) have lower H values than predicted by simple group addition. A specific structural descriptor (*caox-a* in Table 1) was therefore introduced to account for this effect. A similar trend was also observed when the X moiety was located in the β position relative to the carbonyl group and was accounted for by the inclusion of the *caox-b* descriptor. Similarly, H was found to be overestimated for species having a functional group in the α or β position relative to an alcohol moiety. This effect might be linked to some intramolecular H-bonding (e.g., Hine and Mookerjee, 1975; Hilal et al., 2008). This effect was taken into account with the help of two additional structural descriptors: *hyd-a* and *hyd-b*. The inclusion of these 4 descriptors was found to greatly improve H estimates for the multifunctional oxygenated species. However, a bias in predicted H was still found for 2 groups of species: o-nitrophenols and halogenated species bearing a carboxylic acid moiety in the α position. Two additional descriptors (*haloic-a*, *onitrofol* in Table 1) were introduced to correct this bias. This is similar to the correction factors applied in the QSAR method developed by Russell et al. (1992) and in the HENRYWIN method.

The 27 descriptors listed above were all found to be significant for the prediction of H at the 99.9% confidence level (see the p-values in Table 1). However, a small bias was observed in the prediction of H for hydrocarbons and a final descriptor (*nogrp* in Table 1) was included to correct this bias. The computed contribution for *nogrp* remains low and this factor is the least significant in the regression (see the p-value in Table 1).

Figure 2 shows the performance of GROMHE. The scatter plot for the training set in Fig. 2 shows that one species,

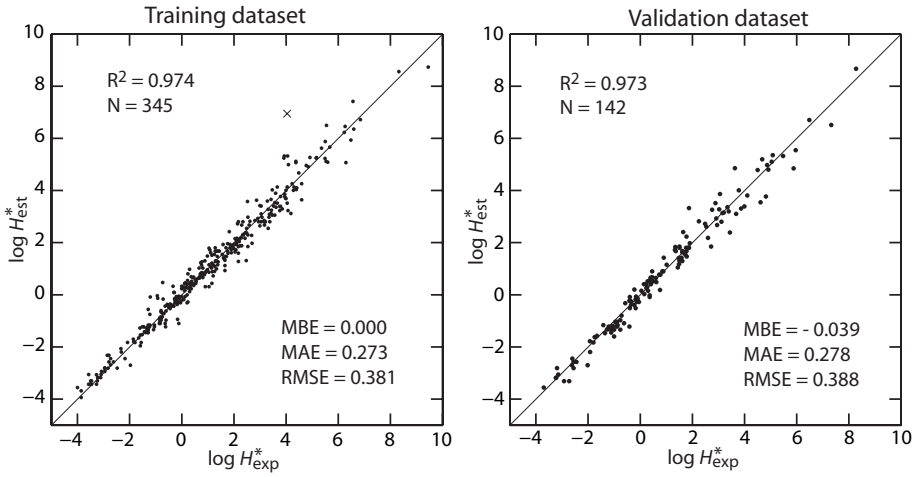

Fig. 2. Scatter plot of estimated $\log H^*$ using GROMHE versus experimental $\log H^*$ for the training set (left panel) and the validation set (right panel). The line is the $y = x$ line. The (x) symbol represents oxo-acetic acid.

oxo-acetic acid, behaves as an outlier with $\log H^*$ overestimated by 3 log units. This species was found to be overestimated by 6 log units using SPARC (see below). The reason of this large overestimation remains unresolved to us and we decided to remove that species from the GROMHE optimisation training set. For the purpose of the intercomparison, oxo-acetic acid was also removed from the statistical analysis. The reliability of the predictions were assessed using the Root Mean Square Error (RMSE), determined as described previously in the context of the hydration constant assessment (see Equation 9), the Mean Absolute Error (MAE) and the Mean Bias Error (MBE):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\log H^*_{\mathrm{est}} - \log H^*_{\mathrm{exp}}\right| \quad (11)$$

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}\left(\log H^*_{\mathrm{est}} - \log H^*_{\mathrm{exp}}\right) \quad (12)$$

where n is the number of species included in the database. The MAE, MBE and RMSE are given in Fig. 2 for the training and validation data sets. Figure 2 shows that the model explains 97% of the total variance of the validation data set. Estimated $\log H^*$ values for the validation set shows no significant bias (MBE = 0.04). The RMSE for the validation set is 0.39, which corresponds to an estimation ability of H^* within a factor of 2.5. The RMSE, MAE, MBE and R^2 values for the validation set are similar to those calculated for the training set and show that the model is not over-fitted (see Fig. 2).

3.3 Analysis of GROMHE estimation method

The previous section shows that GROMHE provides reliable estimates of H^*. The contribution of the descriptors was optimised to obtain a more representative model using the full database. These final contributions agree with those computed for the training data set within their statistical uncertainties (see Table 1). The analysis of GROMHE predictions were performed using the optimised contributions.

The overall performance of GROMHE is summarised in the scatter plot shown in Fig. 3a. The RMSE, MAE and MBE are shown in Fig. 4 together with the box plot of error distribution. The assessment was also performed for different subsets to identify possible bias for various groups of species. Three categories of subsets were defined according to: 1) the number of functional groups (hydrocarbons, monofunctional, difunctional and multifunctional), 2) the aromatic or aliphatic structure of the molecule and 3) the range of the Henry's law constant to differentiate fairly insoluble species (with H^* below $10^3\,\mathrm{Matm}^{-1}$) from more soluble species (H^* greater than $10^3\,\mathrm{Matm}^{-1}$).

The coefficient of determination R^2 between experimental and predicted $\log H^*$ is 0.97 (see Fig. 3a). No significant MBE was found for any of the subsets (see Fig. 4) and thus the GROMHE method seems to provide no systematic bias. Box and scatter plots show that the error increases from simple hydrocarbons to multifunctional species. The RMSE is 0.30 for hydrocarbons and reaches a maximum of 0.52 for difunctional species (see Fig. 4). Similarly, the error in predicting $\log H^*$ for more soluble compounds (i.e. more oxygen substituted compounds) is significantly greater than for less soluble species. The RMSE is 0.33 and 0.53 for the subset of species having H^* below and above $10^3\,\mathrm{Matm}^{-1}$, respectively. It was also observed that the method provides better estimates for the aliphatic subset of species compared to the aromatic subset (see Fig. 4). For the full database, GROMHE finally gives fairly reliable $\log H^*$ estimates, with RMSE of 0.38 and MAE of 0.27.

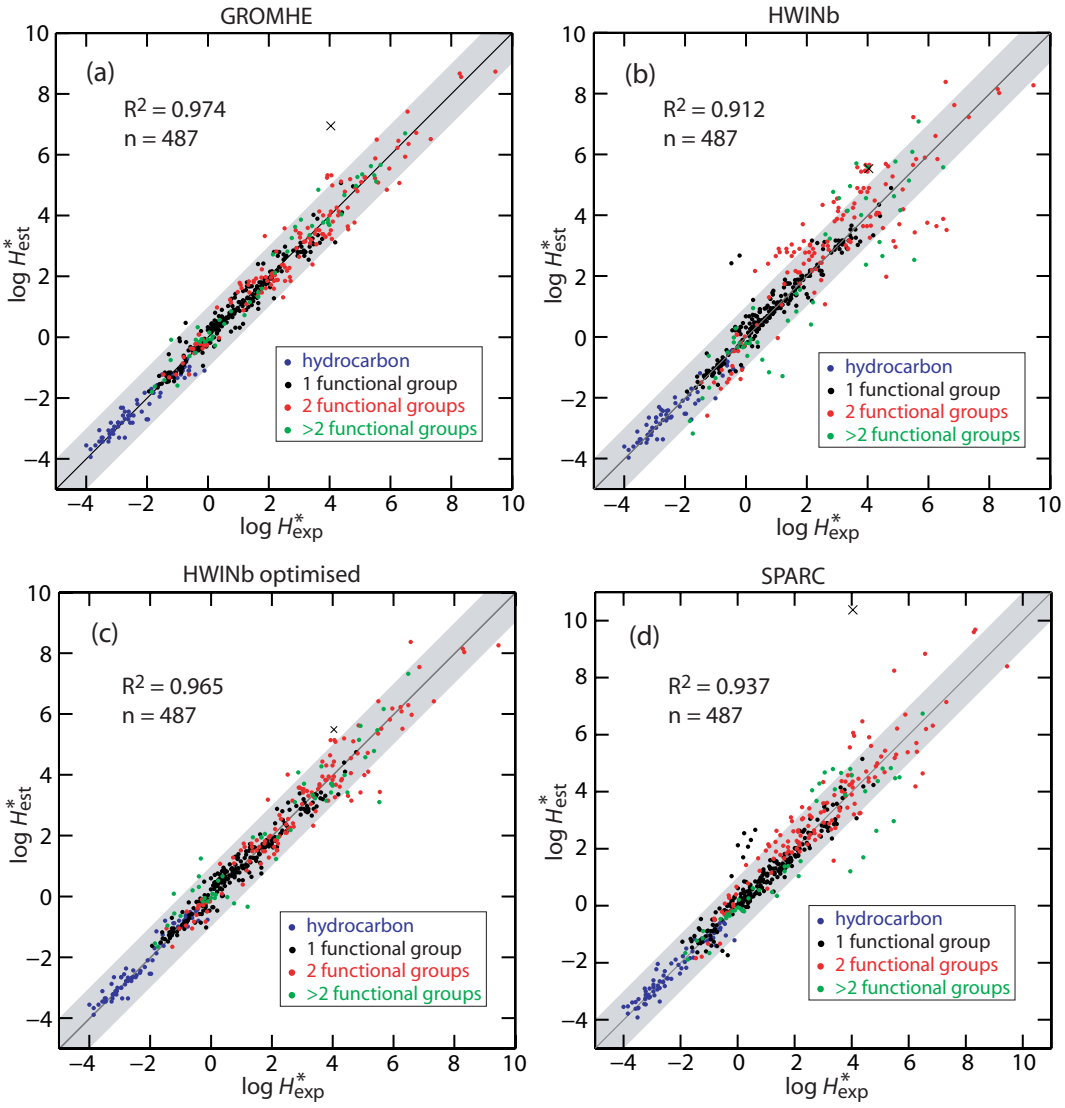

Fig. 3. Scatter plot of estimated versus experimental $\log H^*$ for **(a)** GROMHE method, **(b)** HWINb method, **(c)** HWINb optimised method and **(d)** SPARC-v4.2 method. The line is the $y = x$ line and the grey area represents agreement within one log unit. The (x) symbol represents oxo-acetic acid.

4 Analysis of HWINb and SPARC estimation methods

HWINb, SPARC methods are able to estimate H^* for all the species selected in the database (see Table S1 in the electronic supplement). H^* estimates from HWINb method were determined using the software EPIWIN suit (http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm). The SPARC calculator (http://sparc.chem.uga.edu/sparc) estimates independently the intrinsic H and the hydration constant K_{hyd}. These two properties were jointly used to retrieve the effective Henry's law constant H^* from Eq. (4). The values reported in Table S1 refer to the retrieved $\log H^*$. The overall performance of the models HWINb and SPARC are summarised in scatter plots (Fig. 3b and d). The RMSE, MAE and MBE for each method are shown in Figs. 5 and 6 together with the box plot of error distribution.

For the method HWINb, the scatter plot is shown in Fig. 3b and the performance in Fig. 5. The coefficient of determination for $\log H^*_{est}$ versus $\log H^*_{exp}$ is $R^2 = 0.91$. Hydrocarbon and monofunctional compounds are well predicted with a performance similar to GROMHE's performance. However, their prediction error is much larger for multifunctional compounds with RMSE above 1.0 log unit (see Fig. 5). A bias was also found with a slight tendency towards overestimation of $\log H^*$ for difunctional species and underestimation for species having more than 2 functional groups. This prediction error shows a behaviour similar to that seen for GROMHE, i.e. the error grows with increasing solubility.

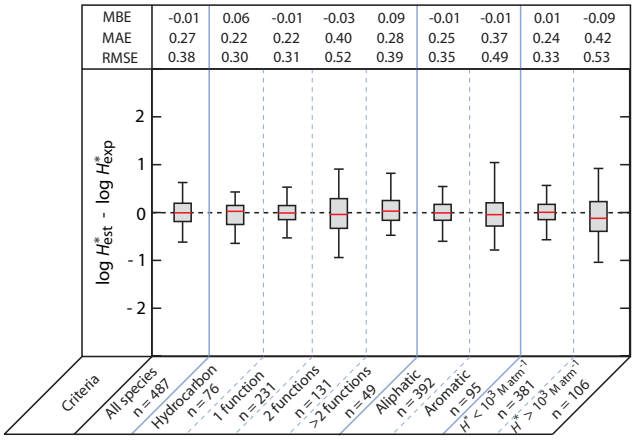

Fig. 4. Mean bias error (MBE), mean absolute error (MAE), root mean square error (RMSE) and box plot for the error distribution in the estimated $\log H^*$ value with the GROMHE method. The whiskers of the box plot show the 5th and 95th percentiles, the box shows the second and third quartile and the red line gives the median value of the distribution.

For the subset of species with H^* above $10^3\,\mathrm{Matm}^{-1}$, the RMSE reaches 1.1 log units which is twice the RMSE given by the GROMHE method. Furthermore, the error obtained with GROMHE for each subset is systematically lower than those obtained using HWINb.

The evaluation of HWINb method was made using the contributions provided by running the EPIWIN software. These contributions are computed using another training database. In an attempt to make a fairer inter-comparison, a multiple linear regression was performed to optimise the HWINb model to our database. The contributions of 47 descriptors (35 bonds and/or fragments and 12 correction factors) were revaluated to describe the structure of the 488 molecules included in the database. The scatter plot obtained with the optimised HWINb model is shown in Fig. 3c and the box plot in Fig. 5. The determination coefficient R^2 is 0.96 compared to 0.91 for the original model. This optimised model shows an improvement especially for the estimation of $\log H^*$ for the more soluble species ($H^* > 10^3\,\mathrm{Matm}^{-1}$) with an RMSE of 0.66 compared to 1.12 for the original model. The MBE was also considerably better showing no bias for all the data subsets. However, the RMSE and the MAE for the subset of species having at least 3 functional groups still remain significantly lower for GROMHE.

The correlation $\log H^*$ estimated using SPARC versus experimental $\log H^*$ is shown in Fig. 3d. The coefficient of determination R^2 is 0.94. SPARC performance is shown in Fig. 6. SPARC and HWINb show similar reliability with similar trends in the prediction of $\log H^*$ for the various subsets. Hydrocarbons and monofunctional compounds are well represented (RMSE < 0.5) whilst errors become large for multifunctional species (RMSE $= 0.97$). Similar to HWINb

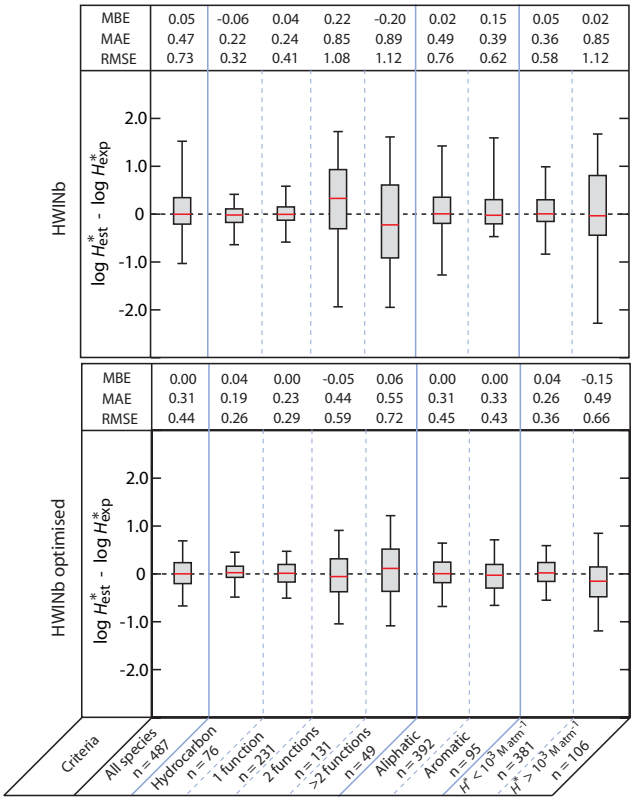

Fig. 5. Mean bias error (MBE), mean absolute error (MAE), root mean square error (RMSE) and box plot for the error distribution in the estimated $\log H^*$ value with the HWINb (top panel) and HWINb optimised method (bottom panel). The whiskers of the box plot show the 5th and 95th percentiles, the box shows the second and third quartile and the red line gives the median value of the distribution.

results, a bias towards $\log H^*$ overestimation is found for difunctional species and towards underestimation for tri or more functional species. Like GROMHE and HWINb the reliability of SPARC estimates decreases with increasing solubility. The error is about one order of magnitude for species having H^* above $10^3\,\mathrm{Matm}^{-1}$ (see Fig. 6). Here again, a fair comparison would require to revaluate the contribution of the SPARC descriptors using our database. However, the SPARC method is based in physical parameters which are determined using quantum mechanical calculations. These calculations are beyond the scope of this paper. Additional work would thus be required to evaluate the inherent performance of the SPARC method before reaching any final conclusions.

5 Conclusions

A new group contribution method, GROMHE, was developed in this study to estimate H^* for organic compounds at 298 K. A multiple linear regression was performed using

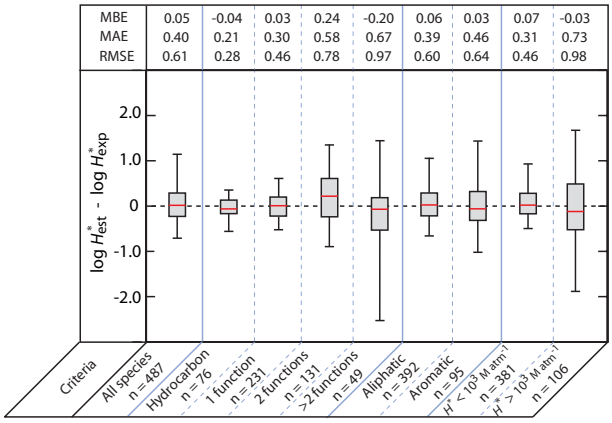

Fig. 6. Mean bias error (MBE), mean absolute error (MAE), root mean square error (RMSE) and box plot for the error distribution in the estimated $\log H^*$ value with the SPARC-v4.2 method. The whiskers of the box plot show the 5th and 95th percentiles, the box shows the second and third quartile and the red line gives the median value of the distribution.

the training data set including 345 organics representative species of atmospheric interest. A set of 28 descriptors was found to be statistically significant for the prediction of $\log H^*$. The resulting method predicts $\log H^*$ with a root mean square error of 0.39 for a validation set including 142 species. No statistically significant bias was observed. The regression fit of predicted versus observed $\log H^*$ shows a coefficient of determination of $R^2 = 0.97$.

The H^* values for hydrocarbon and monofunctional compounds were well predicted with similar performance with all the methods assessed (GROMHE, HWINb and SPARC). The results show that the reliability of the predicted values decreases when H^* increases. Species having H^* above $10^3\,\mathrm{Matm}^{-1}$ are of particular interest in the context of atmospheric chemistry. For the subset of species having H^* above that threshold, the RMSE obtained for GROMHE is 0.53 log units. For the same subset, the reliability of the prediction using HWINb or SPARC was appreciably lower with RMSE of about 1 log unit. However, to some extent this large error can be explained by changing the dataset used for training and validation. In particular, the performance of HWINb was found to be significantly improved when the contribution of the descriptors was recalculated using the data compiled in this work. The results show that GROMHE performs well compared to the other SARs. These results give confidence in the ability of GROMHE to determine the H^* for organics known to be important in atmospheric chemistry.

Supplementary material related to this article is available online at: http://www.atmos-chem-phys.net/10/7643/2010/acp-10-7643-2010-supplement.pdf.

Acknowledgements. We would like to thank S. H. Hilal and L. A. Carreira for providing us with data and support to run SPARC. Helpful comments on the manuscript by A. Dutot and J. Lee-Taylor and Adriana Coman are gratefully acknowledged. Van Viet Ngo and Romain Bouchaud contributed to the early development of GROMHE.

Edited by: M. Petters

The publication of this article is financed by CNRS-INSU.

References

Atkinson, R.: Atmospheric chemistry of VOCs and NO_x, Atmos. Environ., 34, 2063–2101, 2000.

Aumont, B., Madronich, S., Bey, I., and Tyndall, G. S.: Contribution of secondary VOC to the composition of aqueous atmospheric particles: a modeling approach, J. Atmos. Chem., 35, 59–75, 2000.

Aumont, B., Szopa, S., and Madronich, S.: Modelling the evolution of organic carbon during its gas-phase tropospheric oxidation: development of an explicit model based on a self generating approach, Atmos. Chem. Phys., 5, 2497–2517, doi:10.5194/acp-5-2497-2005, 2005.

Betterton, E. A., and Hoffmann, M. R.: Henry's law constants of some environmentally important aldehydes, Environ. Sci. Technol, 22, 1415–1418, 1988.

Blando, J. D., and Turpin, B. J.: Secondary organic aerosol formation in cloud and fog droplets: a literature evaluation of plausibility, Atmos. Environ., 34, 1623–1632, 2000.

Boethling, R. S., and Mackay, D.: Handbook of Property Estimation Methods for Chemicals: Environmental Health Sciences, CRC Press LLC, Washington, DC, USA, 2000.

Camredon, M., Aumont, B., Lee-Taylor, J., and Madronich, S.: The SOA/VOC/NO_x system: an explicit model of secondary organic aerosol formation, Atmos. Chem. Phys., 7, 5599–5610, doi:10.5194/acp-7-5599-2007, 2007.

Chen, J., Griffin, R. J., Grini, A., and Tulet, P.: Modeling secondary organic aerosol formation through cloud processing of organic compounds, Atmos. Chem. Phys., 7, 5343–5355, doi:10.5194/acp-7-5343-2007, 2007.

Dearden, J. C. and Schuurmann, G.: Quantitative structure-property relationships for predicting Henry's law constant from molecular structure, Environ. Toxicol. Chem., 22, 1755–1770, 2003.

Ervens, B., George, C., Williams, J. E., Buxton, G. V., Salmon, G. A., Bydder, M., Wilkinson, F., Dentener, F., Wolke, R., and Herrmann, H.: CAPRAM 2.4 (MODAC mechanism): an extended and condensed tropospheric aqueous phase mechanism and its application, J. Geophys. Res.-Atmos., 108, 4426, doi:4410.1029/2002JD002202, 2003.

Ervens, B., Carlton, A. G., Turpin, B. J., Altieri, K. E., Kreidenweis, S. M., and Feingold, G.: Secondary organic aerosol yields from cloud-processing of isoprene oxidation products, Geophys. Res. Lett., 35, 5, L02816, doi:10.1029/2007JL031828, 2008.

Facchini, M. C., Fuzzi, S., Zappoli, S., Andracchio, A., Gelencser, A., Kiss, G., Krivacsy, Z., Meszaros, E., Hansson, H. C., Alsberg, T., and Zebuhr, Y.: Partitioning of the organic aerosol component between fog droplets and interstitial air, J. Geophys. Res.-Atmos., 104, 26821–26832, 1999.

Finlayson-Pitts, B. J. and Pitts, J. N.: Chemistry of the upper and lower atmosphere, Academic Press, San Diego, USA, 2000.

Gelencser, A., and Varga, Z.: Evaluation of the atmospheric significance of multiphase reactions in atmospheric secondary organic aerosol formation, Atmos. Chem. Phys., 5, 2823–2831, doi:10.5194/acp-5-2823-2005, 2005.

Goldstein, A. H., and Galbally, I. E.: Known and unexplored organic constituents in the earth's atmosphere, Environ. Sci. Technol., 41, 1514–1521, 2007.

Griffin, R. J., Nguyen, K., Dabdub, D., and Seinfeld, J. H.: A coupled hydrophobic-hydrophilic model for predicting secondary organic aerosol formation, J. Atmos. Chem., 44, 171–190, 2003.

Hallquist, M., Wenger, J. C., Baltensperger, U., Rudich, Y., Simpson, D., Claeys, M., Dommen, J., Donahue, N. M., George, C., Goldstein, A. H., Hamilton, J. F., Herrmann, H., Hoffmann, T., Iinuma, Y., Jang, M., Jenkin, M. E., Jimenez, J. L., Kiendler-Scharr, A., Maenhaut, W., McFiggans, G., Mentel, T. F., Monod, A., Prevot, A. S. H., Seinfeld, J. H., Surratt, J. D., Szmigielski, R., and Wildt, J.: The formation, properties and impact of secondary organic aerosol: current and emerging issues, Atmos. Chem. Phys., 9, 5155–5236, doi:10.5194/acp-9-5155-2009, 2009.

Hansch, C., Leo, A. , and Hoekman, D.: Exploring QSAR: Hydrophobic, Electronic, and Steric Constants, American Chemical Society, Washington DC, USA, 1995.

Herrmann, H., Ervens, B., Jacobi, H.-W., Wolke, R., Nowacki, P., and Zellner, R.: CAPRAM2.3: A chemical aqueous phase radical mechanism for tropospheric chemistry, J. Atmos. Chem., 36, 231–284, 2000.

Herrmann, H., Tilgner, A., Barzaghi, P., Majdik, Z., Gligorovski, S., Poulain, L., and Monod, A.: Towards a more detailed description of tropospheric aqueous phase organic chemistry: CAPRAM 3.0, Atmos. Environ., 39, 4351–4363, 2005.

Hilal, S. H., Karickhoff, S. W., and Carreira, L. A.: Prediction of the solubility, activity coefficient and liquid/liquid partition coefficient of organic compounds, Qsar Comb. Sci., 23, 709–720, 2004.

Hilal S. H., Ayyampalayam S. N., and Carreira, L. A.: Air-liquid partition coefficient for a diverse set of organic compounds: Henry's law constant in water and hexadecane, Environ. Sci. Technol., 42(24), 9231–9236, 2008.

Hine, J. and Moorkerjee, P. K.: The intrinsic hydrophilic character of organic compounds, Correlations in terms of structural contributions, J. Org. Chem., 40, 292–298, 1975.

Jacob D., Gottlieb, E. W. and Prather, M. J.: Chemistry of polluted cloudy boundary layer, J. Geophys. Res.-Atmos., 94, 12975–13002, 1989.

Kuhne, R., Ebert, R. U., and Schuurmann, G.: Prediction of the temperature dependency of Henry's law constant from chemical structure, Environ. Sci. Technol., 39(17), 6705–6711, 2005.

Le Henaff, P.: Methodes d'etude et proprietes des hydrates, hemiacetals et hemithioacetals derives des aldehydes et des cetones., P. Bull. Soc. Chim. Fr., 11, 4687–4698, 1968.

Legrand, M., Preunkert, S., Wagenbach, D., Cachier, H., and Puxbaum, H.: A historical record of formate and acetate from a high-elevation alpine glacier: Implications for their natural versus anthropogenic budgets at the European scale, J. Geophys. Res.-Atmos., 108(15), 4788, doi:10.1029/2003JD003594, 2003.

Legrand, M., Preunkert, S., Galy-Lacaux, C., Liousse, C., and Wagenbach, D.: Atmospheric year-round records of dicarboxylic acids and sulfate at three French sites located between 630 and 4360 m elevation, J. Geophys. Res., 110, D13302, doi:10.1029/2004JD005515, 2005.

Lelieveld, J., and Crutzen, P. J.: Influences of cloud photochemical processes on tropospheric ozone, Nature, 343, 227–233, 1990.

Levine I.N.: Physical Chemistry, fifth edition, McGraw-Hill Higher Education, New York, USA, 2002.

Lim, H. J., Carlton, A. G., and Turpin, B. J.: Isoprene forms secondary organic aerosol through cloud processing: Model simulations, Environ. Sci. Technol., 39(12), 4441–4446, 2005.

Lin, S. T., and Sandler, S. I.: Henry's law constant of organic compounds in water from a group contribution model with multipole corrections, Chem. Eng. Sci., 57, 2727–2733, 2002.

Mackay, D., and Shiu, W. Y.: A critical-review of Henrys law constants for chemicals of environmental interest, J. Phys. Chem. Ref. Data., 10, 1175–1199, 1981.

Meylan, W. M. and Howard, P. H.: Bond contribution method for estimating Henry's law constants, Environ. Toxicol. Chem, 10, 1283–1293, 1991.

Meylan, W. M. and Howard, P. H.: Src's epi suite, v3.20, Syracuse Research Corporation: Syracuse. NY, 2000.

Perrin, D. D., Dempsey, B., and Serjeant, E. P.: pKa prediction for organic acids and bases, Chapman and Hall, London, UK and New York, USA, 1981.

Pun, B. K., Griffin, R. J., Seigneur, C., and Seinfeld, J. H.: Secondary organic aerosol-2. Thermodynamic model for gas/particule partitioning of molecular constituents, J. Geophys. Res., 107(D17), 4333, doi:10.1029/2001JD000542, 2002.

Russell, C. J., Dixon, S. L., and Jurs, P. C.: Computer-assisted study of the relationship between molecular-structure and Henry Law constant, Anal. Chem., 64, 1350–1355, 1992.

Sander, R.: Compilation of Henry's law constants for inorganic and organic species of potential importance in environmental chemistry (Version 3): www.henrys-law.org, 1999.

Saunders, S. M., Jenkin, M. E., Derwent, R. G., and Pilling, M. J.: Protocol for the development of the Master Chemical Mechanism, MCM v3 (Part A): tropospheric degradation of non-aromatic volatile organic compounds, Atmos. Chem. Phys., 3, 161–180, doi:10.5194/acp-3-161-2003, 2003.

Saxena, P., and Hildemann, L. M.: Water-soluble organics in atmospheric particles : a critical review of the literature and application of thermodynamics to identify candidate compounds, J. Atmos. Chem., 24, 57–109, 1996.

Seinfeld, J. H. and Pandis, S. N.: Atmospheric chemistry and physics, Wiley, New York, USA, 1997.

Suzuki, T., Ohtaguchi, K., and Koide, K.: Application of principal components-analysis to calculate Henry constant from molecular-structure, Comput. Chem., 16, 41–52, 1992.

Walcek, C. J., Yuan, H. H., and Stockwell, W. R.: The influence of aqueous-phase chemical reactions on ozone formation in polluted and nonpolluted clouds, Atmos. Environ., 31, 1221–1237, 1997.

Yu, L. E., Shulman, M. L., Kopperud, R., and Hildemann, L. M.: Characterization of organic compounds collected during southeastern aerosol and visibility study: Water-soluble organic species, Environ. Sci. Technol., 39(3), 707–715, 2005.