Overview of response to reviewers

Dear Editor. Please find attached a point-by-point overview of changes made to the manuscript as requested by all referees. The uploaded manuscript also highlights where all changes are made. The main points are as follows, with specific location details of changes discussed in the rest of the document:

- In response to Executive Editor A Kerkweg, we have added a version number to the title and also made the entire code open source with a DOI of the code in its present form. These details are now included in the document as detailed in the responses below. A new section entitled 'code availability' at the end re-iterates where the code can be found.
- In response to referee Takahama, we have added a new separate section that discusses the user interface and file formats. In addition we have clarified the guarantee on functional group specificity, both within the manuscript and on the website, including a more detailed discussion on the SMARTS libraries used. We have also now included text files to be used with examples as supplementary material.
- In response to referee Epstein we have linked to existing services and clarified restrictions on the use of the predictive techniques.
- In response to referee 3, we have clarified definitions regarding properties and non-ideality.

In addition to the requests made by reviewers, we have also added a student name (Nick Dingle) to the author list that was erroneously missed in the first iteration. We have also added acknowledgement to Prof Markus Peters for a discussion on the benefits of making the entire suite open source. We feel the new manuscript is much improved as a result of all reviews.

Detailed responses:

Referee, Executive Editor, A Kerkweg

Comment: *In particular, please note that for your paper, the following requirements have not been met in the Discussions paper:*
• "The main paper must give the model name and version number (or other unique identifier) in the title."
Response: This has now been corrected and the title now reads '*UManSysProp v1.0, an online and open-source facility for molecular property prediction and atmospheric aerosol calculations*.'

Comment: *• "All papers must include a section, at the end of the paper, entitled 'Code availability'. Here, either instructions for obtaining the code, or the reasons why the code is not available should be clearly stated. It is preferred for the code to be uploaded as a supplement or to be made available at a data repository with an associated DOI (digital object identifier) for the exact model version described in the paper. Alternatively, for established models, there may be an existing means of accessing the code through a particular system. In this case, there must exist a means of permanently accessing the precise model version described in the paper. In some cases, authors may prefer to put models on their own website, or to act as a point of contact for obtaining the code. Given the impermanence of websites and email addresses, this is not encouraged, and authors should consider improving the availability with a more permanent arrangement. After the paper is accepted the model archive should be updated to include a link to the GMD paper."*

Response: In addition to providing the online portal for users who do not want to use source code, and the JSON API for linking with our web portal without using a web browser, we also now provide the source code for all predictive techniques provided on the site. We have released this via a github repository that has an associated DOI 10.5281/zenodo.45143 as requested by GMD. This information has now been added to the introduction and abstract via the text below, on the web site and in a new section entitled '*Code Availability*' at the end:

'*In addition to providing the online portal for users who do not want to use source code, and the JSON API for linking with our web portal without using a web browser, we also provide the source code for all predictive techniques provided on the site, covered by the GNU GPL license to encourage development of a user community. We have released this via a Github repository (https://github.com/loftytopping/UManSysProp_public.git), that has an associated DOI for the exact*

model version given in this paper as provided by the Zenodo service (DOI 10.5281/zenodo.45143)"

Referee S.A.Epstein

General Comments
***Comment:** . A brief summary of other thermodynamic property prediction facilities (such as EPI Suite) would be a useful addition to the introduction.*
Response: We agree. In the introduction, after the sentence: '*..the development of community driven software at least enables modellers to tackle this problem directly*' we have added the following line. '*There are a number of property predictions facilities that are available online. For example, the US EPA host predictive models and tools for assessing chemicals under the Toxic Substances Control Act (TSCA) (http://www.epa.gov/tsca-screening-tools). From this site one can access the simulation program 'Estimation Programs Interface' (EPI) Suite facility (http://www.epa.gov/tsca-screening-tools/download-epi-suitetm-estimation-program-interface-v411). This provides a number of facilities including estimates of physical / chemical properties (melting point, water solubility, etc.) and environmental fate properties (breakdown in water or air, etc.). The Dortmund Databank (DDB) provide a wide range of database and software products related to fundamental properties of molecules and mixtures. With varying proprietary and free educational access, their program package ARTIST was developed for the estimation of pure component properties from molecular structure. In the UK the National Chemical Database Service (CDS) provides free access to web-based services including ACD/Labs Inc Physchem and NMR predictions (http://cds.rsc.org/). Services specifically tailored to atmospheric studies include the E-AIM community model for calculating gas/solid/liquid partitioning (http://www.aim.env.uea.ac.uk/aim/aim.php) and the AIOMFAC portal for calculating activity coefficients in mixed inorganic/organic liquid systems (http://www.aiomfac.caltech.edu/).*'

Comment: In cases where physical property measurements are available, the thermodynamic calculations that rely on these physical properties would be more accurate if it was possible to use the measurements instead of the predicted values. The ability to use physical property measurement values, when available, in the aerosol calculations will significantly improve the accuracy and utility of the prediction facilities.
Response: We agree. In a follow-up development we now have funds to link the existing website to a new database of measurements linked to property predictions provided. Creating a standardized database in itself can be challenging, as noted in the recent review of saturation vapour pressures by Bilde et al (2015). To allude to these developments, we have added the following sentence to the section 'Future work': *Where property measurements are available, these might prove more accurate than any given estimation technique. With this in mind, in addition to extending the range of predictions provided, UManSysProp will also be linked to a standardized database of property measurements*.'

Bilde et al (2015). Saturation Vapor Pressures and Transition Enthalpies of Low-Volatility Organic Molecules of Atmospheric Relevance: From Dicarboxylic Acids to Complex Mixtures. Chemical Reviews 2015 115 (10), 4115-4156 DOI: 10.1021/cr5005502

Specific Comments
1. Page 9673: It is unclear whether CAS numbers can be used directly in the prediction facility
Response: Presently they cannot. We have added the following line to the table caption:' *Please note, CAS numbers cannot be used directly in the prediction facility.'*

2. Page 9675, lines 21-23: It is unclear what the phase "Pure component properties are limited to 5000 compounds, predictions involving activity coefficients limited to 1000 compounds" means. Are these limits on the number of molecules that can be submitted at one time? **Response:** Yes this is correct. We have limited the number of compounds for different predictions due to computational burden and delays on providing results. We have added the following text to this sentence:'[..] *at any one time via the web portal*'.

3. Page 9676, line 21-22: It would be helpful to have the methodology options for predicting sub-cooled liquid density predictions in this bullet-point, as in the previous bullet-points
Response: This has now been added.

4. Page 9676, line 23-24: It would be helpful to have the methodology options for predicting pure

component vapor pressures in this bullet-point, as in the previous bullet-points
Response: This has now been added.

5. Figure 3: A key to the axis labels should be provided in the caption
Response: This has now been added.

6. Figure 3: It may make more sense to combine both plots and use different marker symbols to represent each vapor pressure prediction
Response: This has been changed.

7. Section 3.2: The following papers should be cited when introducing predictions of absorptive partitioning.

a. Pankow, J. An absorption model of the gas/aerosol partitioning involved in the formation of secondary organic aerosol. Atmos. Environ. 1994, 28, 189.
b. Donahue, N.M., Robinson, A.L., Stanier, C.O., Pandis, S.N., 2006. Coupled partitioning, dilution, and chemical aging of semivolatile organics. Environ. Sci. Technol. 40, 2635–2643.

Response: This has been changed and added to the line 'Equilibrium absorptive partitioning […]' in section 3.

Technical Corrections
1. Page 9676, line 10: "xx" was never replaced with a number
Response: This has now been corrected.

2. Page 9678, line 16: the quantity in parenthesis should be raised to a power of -1
Response: This has now been corrected.

3. Page 9681, line 13: "re-partitioning" should be "re-partition"
Response: This has now been corrected.

4. Page 9705: there should be a space after "size" in the first line of the figure caption
Response: This has now been corrected.

Referee S. Takahama

Comment: *The description of the user interface and file formats for information exchange is mentioned in the main text in Sections 1 and 3, but it would be helpful for readers if a consolidated section is dedicated to its description and kept separate from the properties being predicted. In Appendix A1, it might be helpful to point out to the wide array of potential readers that "any machine" and "other platforms" specifically includes Microsoft Windows and Mac OS X, among other operating systems.*
Response: We agree with this suggestion and have moved relevant details into a new section 3.1, incrementing further sections accordingly. The new section reads as follows

The UManSysProp website first provides a portal where users can enter or upload a SMILES string and predict the property of interest. Examples of supplying SMILES strings via the input are given in section 3.1 and 3.2. Whilst users have the option to display output on a new webpage via HTML as the default option, the following download options are also available. For more information on their use, please refer to the references given in parentheses

- HTML (view in web browser)

- Excel file

- Python pickle file \url{ https://docs.python.org/2/library/pickle.html}

-XML file \url{ https://en.wikipedia.org/wiki/XML}

-Zipped CSV file

-JSON file \url{ https://en.wikipedia.org/wiki/JSON}

If you want to access UManSysProp without using a web-browser we also provide a programmer friendly JSON API that enables you to call our suite of tools from your own code. This is described in detail on our ReadTheDocs.org webpage \url{https://umansysprop.readthedocs.org/} with an example provided in the Appendix where we briefly discuss future expansions.} \\In addition, this section is now referred to in the introduction during a breakdown of the manuscript and specific details moved to new section 3.1.We have also added the point regarding Windows and Mac OSX via the following sentence in section 6.2 \textit{'*[..] provided you have Python 2.7 or greater. This includes Microsoft Windows, Mac OS X and other operating systems*.'

Comment: *The reason for separation from Figure 2 from Figure 1 is unclear, as activity coefficients are bulk parameters (and therefore can be included in Figure 1) and SMARTS patterns are also used for property prediction (and therefore the requirement of SMARTS is not unique to the task depicted in Figure 2). It seems the two figures can be merged to give an overall picture of UManSysProp.*
Response: This has now been changed and figure 2 removed from the document, including reference to it in section 2.1.

Comment: *It would be nice to see the authors expound upon the statement (Section 3, p. 9764): C3383 "For techniques used in UManSysProp, an extensive manual analysis of compounds used in the MCM (Jenkin et al., 2012), and a subset of GECKO mechanism (Aumont et al., 2005), were used to validate derived SMARTS libraries." As the authors note, the formulation of SMARTS patterns require special care to target specificity. Having realized these difficulties in our own work \url{http://www.atmos-chem phys- discuss.net/15/33631/2015/}, we have dedicated a technical note to present methods for validation and the structure of compounds and range (number of groups per compound) for which our formulated patterns were tested. While exhaustive description of the validation process is clearly not the focus of this work, many of the results rest on the correct enumeration of functional groups in molecules. Therefore, further discussion of the validation that the authors have already conducted can benefit users of the software and those seeking to adopt chemoinformatic approaches for structural queries. Additionally, some of the example property estimates presented in the current manuscript to some extent serve as even further validation.*
Response: This is an important point and the referee is quite right in raising it. It can be easy for developers to generate generic SMILES and introduce errors with regards to specificity. In the text that follows the brief introduction to this challenge, we discuss the specific approaches for ensuring certain predictive techniques captured the correct structural features for compounds of atmospheric interest from chemical mechanisms. This was actually carried out manually in the first instance. The facility does record atomic information and uses this to check there are no over- and under-counting of individual atoms by the construction of required functional groups. However we do not have a more detailed mechanism in place to check this for compounds outside of the chemical mechanisms mentioned. To make sure the reader is aware of this, we suggest the following addition to the end of section 2.1 \textit{'*All of the above checks of specificity were carried out by hand for atmospheric chemical mechanisms. Whilst the current facilities check for under- or over-counting of atoms for any given set of functional groups, a future development would need an automatic method of checking specificity for compounds falling outside of this subset following the discussions presented by Ruggeri and Takahama (2015)*.'

Ruggeri, G. and Takahama, S.: Technical Note: Development of chemoinformatic tools to enumerate functional groups in molecules for organic aerosol characterization, Atmos. Chem. Phys. Discuss., 15, 33631-33674, doi:10.5194/acpd-15-33631-2015, 2015.

Comment: *Section 2, Parsing: introduction of examples containing contrasting primary -OH structures and how each are matched by one of the five groups defined might illustrate the elegance of this*

approach better to interested readers.
Response: We have now removed the lines '*As noted in that paper, whilst it is 'easy' to identify all primary alcohols (SMARTS 'a' in the table), the Nannoolal method requires primary alcohols to be split between NG 35 (carbon chain of 5 or more atoms with nomenclature defined in Nannoolal et al. (2008)) and NG 36 (primary alcohols on a C4 or smaller chain) although the exact criteria for this split is not clear in the literature. For our applications, the allocation of primary alcohols is achieved using a set of five SMARTS.*'} and added the following text to illustrate this, taken from the Barley et al (2011) paper: \textit{'*It is easy enough to identify all primary alcohols (SMARTS a in Table 3) but the Nannoolal method requires primary alcohols to be split between NG_35 (carbon chain of 5 or more atoms) and NG_36 (primary alcohols on a C4 or smaller chain) although the exact criteria for this split is not clear in the literature. In our work (Barley et al 2011) the allocation of primary alcohols is achieved using a set of five SMARTS. SMARTS b in Table 3 identifies whether the primary alcohol is on a "carbon" chain of 5 or more atoms. This chain has to be terminated by carbon atoms (which may bear functional groups that are not part of this count), but the intermediate atoms can be N or O as well as C. Hence (using SMILES notation) OCCCO and OCCCCO would both have two alcohol groups belonging to NG_36 while OCCCCC, OCCOCC and OCCN(C)CC would have primary alcohols belonging to NG_35. The other three SMARTS account for the possible branching of this heavy atom chain:- thus OCC(C)(C) and OCN(C)C would both be NG_36 alcohols while OCC(C)(C)C and OCN(C)CC would be NG_35 alcohols.*'
We have now also released all of the source code behind UManSysProp, despite including the web-portal and JSON API, so users can study SMARTS for all methods. By using the OpenBabel framework, we would encourage the user community to cycle through compounds of interest.

Minor comments:

Comment: *The tables in appendices A1-A3 would be better served if provided (also) as text files in supporting information, as successful copying tabular data from PDFs can be depending on PDF viewer. Regarding the web interface, for future versions it may be helpful to include the possibility of providing estimates of pure component properties (e.g., vapor pressures) from multiple methods in the same output file. It is likely that one of the uses for this technique will be to assess uncertainties in estimated properties and such a feature will allow users to more directly answer this question.*
Response: We now provide these tables as text files in the supporting information. We agree on changing output options for future use and will respond to any similar requests during the initial uptake.

Comment: p. 9671, line 18: "It isn't clear" -> "It is not clear" C3384
Response: This has now been corrected.

Comment: p. 9684, line 10: 'by appending ?' -> 'by appending "?"' or 'by appending a question mark'
Response: This has now been corrected.

Comment: p. 9675, line 5: "Fig. 2" -> "Figure 2" p. 9676, line 10: "xx species"
Response: This has now been corrected.

Comment: p. 9678, line 8: "Sub cooled liquid" -> "Sub-cooled liquid"
Response: This has now been corrected.

Comment: p. 9676, line 27: "parenthese:" -> "parentheses:"
Response: This has now been corrected.

Referee 3

Comment:*My main concern is that neither the manuscript nor the website clearly describes the basic conditions and assumptions of the models used for calculating the different properties, which puts the otherwise great tool in danger of becoming a completely black box. I appreciate that the authors refer to previous papers describing model development and validation and that validation against data and other predictive tools is outside the scope of the present paper. But I would as a minimum prefer to have a short*

list of the most fundamental assumptions behind each model and a brief mentioning of cases where other existing models yield significantly different predictions for the same properties.
Response: With regards to the first point, we apologies for this perceived lack of clarity. To counteract this we have placed an extra paragraph at the header of every page to more clearly define conditions covered in the simulations, that are not explicitly covered by the maximum/minimum range of conditions allowed.
For all pure component predictions we have added the following: *All group contribution techniques parse the entered compound, represented by a SMILES string, into specific functional groups according to a SMARTS library. These functional groups combinations are then combined with 'interaction' parameters and/or physical parameters to arrive at a given property estimate according to the underlying equations*.
For all activity coefficient predictions we have added the following: *Activity coefficients are predicted assuming a homogeneous bulk representation, allowing all compounds to interact according to the technique applied. No partitioning between the liquid and another phase is accounted for.*
For all absorptive partitioning simulations we have added the following*:'Absorptive partitioning simulations do not account for any gas phase reactions or gas phase non-ideality, the former not expected to occur under ambient conditions. It is assumed the SMILEs strings and abundances uploaded represent a given point in time of interest to the user.*

Comment:*For example, how is ideality defined, what does it cover? Are all components always assumed either ideal or non-ideal (to whatever degree that may be)?*
Response: For non-ideal simulations, all compounds are accounted for in the calculations. Similarly, for ideal simulations all compounds are ideal.

Comment:*What are the dimensions of output variables?*
Response: For activity coefficients these are dimensionless variables. Nonetheless, we have re-iterated on the header for that page that these calculations refer to the liquid state.

Comment:*Are there any limits on applicability ranges of mixtures?*
Response: There are two aspects to this. Firstly, as we state in the manuscript, aside from referencing detailed evaluation studies in the literature, the purpose of this facility is to enable such investigations. Secondly however, the guarantee of functional group specificity is dependent on evaluating how the SMARTS libraries perform for a range of compound structures. As in response to referee Takahama, and to now clarify this for all users, we will a new section on the 'provenance' page of our site: ***Property prediction specificity:*** *Checks of specificity for any given property predictive technique were carried out by hand for atmospheric chemical mechanisms. Whilst the current facilities check for under or over counting of atoms for any given set of functional groups, a future development would need an automatic method of checking specificity for compounds falling outside of this subset such as those proposed by Ruggeri Takahama (2015). More detail on the quality checks in place for ensuring structural features are captured are described in our paper.*

Comment:*Which components are assumed to interact and what are the assumptions regarding these interactions? For example, are some components in the mixtures always ideal or always inert?*
What is assumed regarding the gas-phase ideality?
Response: Only condensed phase components interact and we assume gas phase ideality (i.e. fugacity = mole fraction) and have now indicated this in new page headers, as detailed above.

Comment:*What are the most basic assumptions behind calculations of pure component vapor pressures?*
Response: The definition of a pure component equilibrium vapour pressure is the pressure exerted by a vapour in thermodynamic equilibrium with its condensed phases at a given temperature in a closed system, and we work to this definition. We rely on the group contribution concept in all current estimation techniques. As noted in the paper, and now added to each page header for such techniques, '*All group contribution techniques parse the entered compound, represented by a SMILES string, into specific functional groups according to a SMARTS library. These functional groups combinations are then combined with 'interaction' parameters and/or physical parameters to arrive at a given property estimate according to the underlying equations*.'

Comment:*Exactly what type of vapor pressures are yielded for e.g. organics over aqueous mixtures, e.g. are they equilibrium partial pressures? This would allow the user to more readily gauge the applicability of the present predictions for their own purposes. It would also prevent obscuring the great complexity and many remaining unresolved aspects and mutual model inconsistencies behind the smooth delivery of*

variables with the present online tool, in particular for new users who do not have extensive experience with aerosol and liquid phase thermodynamic calculations.
Response: Yes this is correct. Where vapour pressures are calculated over mixtures, they are specifically noted as equilibrium vapour pressures above the solution**.**

Comment:*It would be helpful to specify very clearly when a "liquid phase" is an organic mixture or an aqueous phase comprising organics, e.g. p. 9676 l. 7. Or if both options are possible in all cases. In general, the conditions for how water is accounted for would be crucial to specify explicitly, see above.*
Response: We can clarify this presuming the question is directed for any simulations that calculated non-ideality in solution. For each of these simulations, we have added the following text to relevant pages: *As requested, water is included as a separate variable within the 'organic' lists. We do not require you to specify whether the mixture is defined as 'organic' or 'aqueous', rather the predictions of activity coefficients rely solely on the relative concentrations of each component without need for defining a specific composition as reference state*.

Comment: *On a minor note, I was a bit puzzled by the use of the term "Kappa Kohler" values, C3392 e.g. p. 9681. Is this convention replacing the use of symbol K(kappa) and the term Köhler theory?*
Response: Whilst we have targeted, largely, the atmospheric community, there is a chance that users would be confused by simply providing the symbol K(kappa). Rather, we decided to fully reference the original papers from which the variable is defined. For clarity however, we have now replaced the above term with K(kappa) Köhler values.

Comment:*The website requires very different input variable formats, e.g. concentrations scales are requested as molecules, grams, or micromoles per some volume. It could be helpful if it would be possible for the user to give these in a number of different preselected dimensions and then the online tool would make the appropriate conversions for the models.*
Response: This is a good suggestion. Whilst we have used standard units in all linked publications prior to this facility, we will judge user feedback as to whether this would be a useful component in a future release.

Specific comments

Comment:*p. 9676 l. 10: XX species?*
Response: Apologies, this has now been corrected

Comment: *In Sect. 3, it is unclear what it means that properties are limited to 5000 or 1000 compounds. Are these the total number of compounds possible to handle, or at a given time?*
Response: Apologies, this has now been clarified. As is response to referee Epstein, we have limited the number of compounds for different predictions due to computational burden and delays on providing results. We have added the following text to page 9675, lines 21-23:' [..] *at any one time via the web portal*'.

Comment: *In Sect. 3.1, and Figs. 5 and 7, it is unclear what the dimension of vapor pressure is.*
Response: This has now been corrected, the units are Log10 atmospheres.

Comment:*In Fig. 5, activity coefficients are given as unitless with no reference to concentration scale or reference state. The dimension is specified on the website, but this information should be clearly stated in the documentation.*
Response: This has now been added to the figure caption.

Comment: *In Fig. 7, the dimension of mass increase is not specified.*
Response: This has now been added to the figure caption.

Comment: In Fig. 4, I suggest using units of [g cm-3] instead of [g/cc].
Response: This has now been added to the figure caption.

Comment: In Table 6, specify that "dry size" is diameter (if that is the case).

Response: This has now been added to the table caption.