Journal of Quantitative Spectroscopy & Radiative Transfer 130 (2013) 81–99

Contents lists available at ScienceDirect

Journal of Quantitative Spectroscopy & Radiative Transfer

journal homepage: www.elsevier.com/locate/jqsrt

# Verification of the $H_2O$ linelists with theoretically developed tools

Q. Ma [a,*], R.H. Tipping [b], N.N. Lavrentieva [c], A.S. Dudaryonok [c]

[a] NASA/Goddard Institute for Space Studies and Department of Applied Physics and Applied Mathematics, Columbia University, 2880 Broadway, New York, NY 10025, USA
[b] Department of Physics and Astronomy, University of Alabama, Tuscaloosa, AL 35487-0324, USA
[c] V. E. Zuev Institute of Atmospheric Optics SB RAS, 1, Akademician Zuev Square, Tomsk 634021, Russia

ARTICLE INFO

*Article history:*
Received 27 February 2013
Received in revised form
9 July 2013
Accepted 15 July 2013
Available online 26 July 2013

*Keywords:*
HITRAN
Properties of the energy levels and wave functions of $H_2O$ states
The pair identity and the smooth variation rules

ABSTRACT

Two basic rules (i.e., the pair identity and the smooth variation rules) resulting from the properties of the energy levels and wave functions of $H_2O$ states govern how the spectroscopic parameters vary with the $H_2O$ lines within the individually defined groups of lines. With these rules, for those lines involving high $j$ states in the same groups, variations of all their spectroscopic parameters (i.e., the transition frequency, intensity, pressure broadened half-width, pressure-induced shift, and temperature exponent) can be well monitored. Thus, the rules can serve as simple and effective tools to screen the $H_2O$ spectroscopic data listed in the HITRAN database and verify the latter's accuracies. By checking violations of the rules occurring among the data within the individual groups, possible errors can be picked up and also possible missing lines in the linelist whose intensities are above the threshold can be identified. We have used these rules to check the accuracies of the spectroscopic parameters and the completeness of the linelists for several important $H_2O$ vibrational bands. Based on our results, the accuracy of the line frequencies in HITRAN 2008 is consistent. For the line intensity, we have found that there are a substantial number of lines whose intensity values are questionable. With respect to other parameters, many mistakes have been found. The above claims are consistent with a well known fact that values of these parameters in HITRAN contain larger uncertainties. Furthermore, supplements of the missing line list consisting of line assignments and positions can be developed from the screening results.

© 2013 Elsevier Ltd. All rights reserved.

## 1. Introduction

Reliable information retrieval using space and ground-based instruments, and reliable climate and atmosphere modeling require an accurate spectroscopic database such as HITRAN [1–3]. This widely-used database has spectroscopic parameters for the most important molecules in bands from the microwave to the ultraviolet spectral regions. It is obvious that the accuracy of these data is essential for users' applications. This is especially true in accurate atmospheric retrievals involving the very important water vapor molecule. In order to meet the accuracy requirement for the $H_2O$ molecule, the $H_2O$ database in HITRAN has been updated several times [1–3]. The current version is HITRAN $H_2O$ 2008 [3] and the new version will be available soon.

Among all the spectroscopic parameters, the line positions in HITRAN have the highest accuracy because the energy levels of $H_2O$ states in different vibrational modes can be accurately calculated and the line positions are simply their differences. In addition, to check accuracies of the theoretically calculated line positions from

* Corresponding author. Tel.: +1 212 678 5574.
*E-mail addresses:* qma@giss.nasa.gov, qm2@columbia.edu (Q. Ma).

0022-4073/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jqsrt.2013.07.017

experiments is relatively simpler. For the line intensity, its theoretical calculations or experimental determinations become more complicated. As a result, the intensity values in HITRAN contain higher uncertainties than the line positions have. However, in general, their accuracies can still be considered as pretty good.

For other remaining parameters, the stories are quite different. For the pressure broadened half-widths, the temperature exponent, and pressure induced shift, to perform their reliable measurements are much more difficult than the line position and the intensity. Thus, their measured data unavoidably contain larger uncertainties. In addition, given the fact that there are a huge number lines and the ambient temperatures in different parts of the atmosphere are much below room temperature, to provide these data from experimental measurements is not realistic. In fact, there is only a small part of their HITRAN $H_2O$ data provided by the measurements and these values may come from several labs with different set-ups that could introduce extra uncertainties. On the other hand, with respect to theoretical calculations, there are several theoretical formalisms available with which one can calculate these parameters. However, the reliability of calculated values from these formalisms significantly suffers by introducing approximations and assumptions necessary in order to make the calculations feasible. As a result, none of these formalisms can yield satisfactory results that meet the accuracy requirements in most of the applications. At present, the most important one is the Robert–Bonamy formalism [4] because it has been used to derive values of the broadened half-width, induced shift, and temperature exponent of $H_2O$ lines adopted for recent editions of HITRAN [5–7]. As another source, semi-empirical and empirical methods have been developed and they are the simplest way to provide required data [8]. By checking the HITRAN source codes [1,3], it can be confirmed that many of the data in HITRAN do come from this kind of sources. However, the reliability of the predicted values from these methods is questionable because in developing these semi-empirical and empirical models, there are not enough measured data with high qualities available. Besides, there is lack of sound theoretical guidance to provide necessary help. In summary, the current uncertainty for the half-widths, induced shifts, and temperature exponents of $H_2O$ lines, especially those involving high $j$ states in the database cot meet the desired accuracy in users' atmospheric applications. It seems that this situation could last for a while.

Of course, one cannot wait until these challenges are completely fulfilled. One needs to find realistic ways to make progress. In the present study, we provide a simple method to improve the accuracies of the spectroscopic parameters of $H_2O$ lines. Without relying on complicated theories and carrying out time consuming calculations, one can make significant progress by identifying wrong values of the spectroscopic parameters listed in HITRAN $H_2O$ 2008 and providing their alternatives. The resource required is nothing but the current database itself. Meanwhile, the tools used are the pair identity and smooth variation rules discovered and tested in our previous studies [9,10]. The power of the rules is that no matter whether they come from measurements, theoretical calculations, or empirical formulas, all the spectroscopic parameters of $H_2O$ lines involving high $j$ states must follow these rules. Then, based on realistic estimations of the uncertainties associated with these parameters in HITRAN, we conclude that in general, with these two rules, one is able to identify errors existing in the database and to provide their alternatives. Meanwhile, for the pressure broadened half-widths, pressure induced shifts, and the temperature exponents, by smoothing their values in individual groups, more accurate values can be obtained.

In Section 2, we briefly describe the two rules and their physics foundations. Then, we show in detail how to categorize $H_2O$ lines into individually defined groups in Section 3 such that their members' spectroscopy parameters are governed by the two rules. This section contains the most useful information for readers who are interested in applying this method in their works. In Sections 4 and 5, we show how to apply the two rules to improve the accuracy of the HITRAN database. The discussions and conclusions of the present study are given in Section 6. Besides, as a reference, we provide a supplement of the missing line list for the pure rotation band in HITRAN $H_2O$ 2008.

## 2. Properties of $H_2O$ states and the rules established for $H_2O$ lines

In our previous works [9,10], we have discovered that among $H_2O$ states in individually defined groups, their energy levels and wave functions share identity and similarity features. For example, for groups consisting of paired $H_2O$ states whose expressions are given by $j_{j-\Delta,\Delta}$ and $j_{j-\Delta,\Delta+1}$ with $\Delta=0,1,\ldots$, the pairs with high $j$ values have almost the same energy levels and the same absolute values of the wave function coefficients. Meanwhile, both patterns of their energy levels and patterns of their wave functions coefficients vary smoothly as the quantum number $j$ of the pairs varies within the groups. The above conclusions are also valid for groups consisting of paired states given by $j_{\Delta,j-\Delta}$ and $j_{\Delta+1,j-\Delta}$ with $\Delta=0,1,\ldots$ . It turns out that by fully exploiting these properties of the energy levels and the wave functions, one can obtain fruitful benefits.

Based on quantum mechanics, one can claim that $H_2O$ lines of interest are completely described by energy levels and wave functions associated with their initial and final $H_2O$ states. As a result, for all the spectroscopic parameters of $H_2O$ lines, the energy levels and the wave functions of $H_2O$ states associated with these lines are the only sources responsible for causing why and governing how these parameters vary with the line of interest within individually defined groups of lines. It is this fact that enables one to establish the pair identity and the smooth variation rules governing variations of the spectroscopic parameter. In our previous work [9], we have relied on the black theory to establish the two rules. Besides, with theoretically analyzing processes in calculating pressure broadened half-widths and induced shifts, we have used a different approach to verify these two rules again [10]. The two rules state that for two paired lines in the same group, values of their spectroscopic parameters must be almost identical and these values must vary smoothly as

the quantum number $j$ of the pairs varies. However, one has to keep in mind that because the identity and similarity features of the energy levels and the wave functions of $H_2O$ states are not completely perfect, the rules hold for $H_2O$ lines involving high $j$ states within small tolerances. The higher the $j$, the firmer the two rules hold. For each of the groups, one can introduce the boundary $j_{bd}$ as numerical measures such that the pair identity rule should be applicable for pairs with $j \geq j_{bd}$. With respect to the smooth variation rule, it becomes valid significantly earlier than $j_{bd}$ used to justify the applicability for the first rule. Readers can find out how to estimate $j_{bd}$ for specified groups together with suggested boundary values in our previous works [9,10].

## 3. Categorizations of $H_2O$ lines

For later convenience, we label $H_2O$ states as the A type of state or the C type of state by comparing their $k_a$ and $k_c$ values. States labeled as the A type are those whose $k_a$ values are closer to $j$ and states as the C type are those whose $k_c$ values are closer to $j$. For states whose $k_a$ and $k_c$ values become comparative, there is a bias to label them as the A type because the $H_2O$ molecule is closer to a prolate top than an oblate top. Without losing generality, we prefer to assign $k_a = j - \Delta$ and $k_c = \Delta$ or $\Delta + 1$ for the A type states and to assign $k_c = j - \Delta$ and $k_a = \Delta$ or $\Delta + 1$ for the C type states where $\Delta = 0,1,\ldots$ .

It has been shown that within certain sets of the $H_2O$ states, there are the pair identity and smooth variation properties of their energy levels and wave functions. As a result, one should know how to categorize the states accordingly in order to maintain these properties. This knowledge provides a useful hint for how to divide the $H_2O$ lines into groups. In fact, what one needs to do in dividing lines is nothing but to categorize the initial and final states of lines simultaneously. First of all, one divides lines into the P, Q, and R branches. Then within the same branch, one distinguishes the lines into four cases according to which type (i.e., A or C) their initial and final states belong to. Finally, one categorizes lines within the same branches and within the same cases. After the categorization procedure is completed, all those lines whose initial and final states have similar properties are grouped into the same categories.

In the following, we show the categorization results in detail. First of all, we present general expressions for categorized lines in the 000–000, 010–000, and 100–000 bands in Table 1. For example, the categorized lines of the R branch in the A–A case can be expressed in pairs of lines as $j'_{j'-\Delta',\Delta'+1} \leftarrow j''_{j''-\Delta'',\Delta''}$ and $j'_{j'-\Delta',\Delta'} \leftarrow j''_{j''-\Delta'',\Delta''+1}$, where $\Delta' = 0,1,2,\ldots$ and $\Delta'' = 0,1,2,\ldots$ . Although the parameters $\Delta'$ and $\Delta''$ run independently, there are restrictions of their parities applied for the A–A and C–C cases. For lines in the P and R branches, $\Delta'$ and $\Delta''$ must have the same evenness or oddness and for lines in the Q branch, $\Delta'$ and $\Delta''$ must have the opposite evenness or oddness. In contrast, there are no restrictions for the A–C and C–A cases.

It is worth mentioning that due to the selection rule, not all of the pairs listed in Table 1 exist. In fact, the two paired lines existing are those pairs in the A–A and C–C cases. In contrast, in the A–C and C–A cases, due to the selection rule there is only one transition allowed for each partner of the pairs and these allowed transitions are always switched between the two paired lines as the angular momentum $j''$ changes by every one number. We note that the categorization formulas listed in Table 1 are also applicable for other bands whose selection rules are given by $\Delta k_a = \pm 1, \pm 3,\ldots$ and $\Delta k_c = \pm 1, \pm 3,\ldots$ . Finally, there are no lines in the C–A case for the pure rotational band because energy levels of the states with the A type are always higher than those of the states with the C type there. For other bands, there is no such restriction.

It is well known that there are some $H_2O$ bands whose selection rules are not the same as those bands mentioned above. For example, the 001–000 band associated with the antisymmetric stretching motion. The latter causes an oscillating dipole moment along the $a$-axis in the molecular fixed frame. As a result, the selection rules become $\Delta k_a = 0, \pm 2,\ldots$ and $\Delta k_c = \pm 1, \pm 3,\ldots$ . For those bands, the categorizations have different formulas which are listed in Table 2. It is worth mentioning that except for the A–A case, these two kinds of bands share the same groups. For the A–A case, their groups are not overlapped because their restrictions of the evenness and the oddness for $\Delta'$ and $\Delta''$ are always opposite.

In practice, by using the general expressions given above, one may not be able to categorize lines exclusively. In other words, some lines, especially those whose $k_a$ and $k_c$ values are comparable, may fit into more than one expression.

**Table 1**
Categorizations of lines in the 000–000, 010–000, and 100–000 bands.

| Cases | The P and R branches | The Q branch |
|---|---|---|
| A–A | $j'_{j'-\Delta',\Delta'+1} \leftarrow j''_{j''-\Delta'',\Delta''}$<br>$j'_{j'-\Delta',\Delta'} \leftarrow j''_{j''-\Delta'',\Delta''+1}$<br>($\Delta'$ and $\Delta''$ in the same parity) | $j'_{j'-\Delta',\Delta'} \leftarrow j''_{j''-\Delta'',\Delta''}$<br>$j'_{j'-\Delta',\Delta'+1} \leftarrow j''_{j''-\Delta'',\Delta''+1}$<br>($\Delta'$ and $\Delta''$ in the opposite parity) |
| A–C | $j'_{j'-\Delta',\Delta'+1} \leftarrow j''_{\Delta'',j''-\Delta''}$<br>$j'_{j'-\Delta',\Delta'} \leftarrow j''_{\Delta''+1,j''-\Delta''}$ | $j'_{j'-\Delta',\Delta'} \leftarrow j''_{\Delta'',j''-\Delta''}$<br>$j'_{j'-\Delta',\Delta'+1} \leftarrow j''_{\Delta''+1,j''-\Delta''}$ |
| C–C | $j'_{\Delta'+1,j'-\Delta'} \leftarrow j''_{\Delta'',j''-\Delta''}$<br>$j'_{\Delta',j'-\Delta'} \leftarrow j''_{\Delta''+1,j''-\Delta''}$<br>($\Delta'$ and $\Delta''$ in the same parity) | $j'_{\Delta',j'-\Delta'} \leftarrow j''_{\Delta'',j''-\Delta''}$<br>$j'_{\Delta'+1,j'-\Delta'} \leftarrow j''_{\Delta''+1,j''-\Delta''}$<br>($\Delta'$ and $\Delta''$ in the opposite parity) |
| C–A | $j'_{\Delta'+1,j'-\Delta'} \leftarrow j''_{j''-\Delta'',\Delta''}$<br>$j'_{\Delta',j'-\Delta'} \leftarrow j''_{j''-\Delta'',\Delta''+1}$ | $j'_{\Delta',j'-\Delta'} \leftarrow j''_{j''-\Delta'',\Delta''}$<br>$j'_{\Delta'+1,j'-\Delta'} \leftarrow j''_{j''-\Delta'',\Delta''+1}$ |

**Table 2**
Categorizations of lines in the 001–000 and 011–000 bands.

| Cases | The P and R branches | The Q branch |
|---|---|---|
| A–A | $j'_{j'-\Delta',\Delta'} \leftarrow j''_{j''-\Delta'',\Delta''}$<br>$j'_{j'-\Delta',\Delta'+1} \leftarrow j''_{j''-\Delta'',\Delta''+1}$<br>($\Delta'$ and $\Delta''$ in the opposite parity) | $j'_{j'-\Delta',\Delta'+1} \leftarrow j''_{j''-\Delta'',\Delta''}$<br>$j'_{j'-\Delta',\Delta'} \leftarrow j''_{j''-\Delta'',\Delta''+1}$<br>($\Delta'$ and $\Delta''$ in the same parity) |
| A–C | $j'_{j'-\Delta',\Delta'} \leftarrow j''_{\Delta'',j''-\Delta''}$<br>$j'_{j'-\Delta',\Delta'+1} \leftarrow j''_{\Delta''+1,j''-\Delta''}$ | $j'_{j'-\Delta',\Delta'+1} \leftarrow j''_{\Delta'',j''-\Delta''}$<br>$j'_{j'-\Delta',\Delta'} \leftarrow j''_{\Delta''+1,j''-\Delta''}$ |
| C–C | $j'_{\Delta',j'-\Delta'} \leftarrow j''_{\Delta'',j''-\Delta''}$<br>$j'_{\Delta'+1,j'-\Delta'} \leftarrow j''_{\Delta''+1,j''-\Delta''}$<br>($\Delta'$ and $\Delta''$ in the same parity) | $j'_{\Delta'+1,j'-\Delta'} \leftarrow j''_{\Delta'',j''-\Delta''}$<br>$j'_{\Delta',j'-\Delta'} \leftarrow j''_{\Delta''+1,j''-\Delta''}$<br>($\Delta'$ and $\Delta''$ in the opposite parity) |
| C–A | $j'_{\Delta',j'-\Delta'} \leftarrow j''_{j''-\Delta'',\Delta''}$<br>$j'_{\Delta'+1,j'-\Delta'} \leftarrow j''_{j''-\Delta'',\Delta''+1}$ | $j'_{\Delta'+1,j'-\Delta'} \leftarrow j''_{j''-\Delta'',\Delta''}$<br>$j'_{\Delta',j'-\Delta'} \leftarrow j''_{j''-\Delta'',\Delta''+1}$ |

**Table 3**
Numbers of lines and groups in the P, Q, and R branches for several bands.

| Bands | Total # of lines | The P branch | | The Q branch | | The R branch | |
|---|---|---|---|---|---|---|---|
| | | # lines | # groups | # lines | # groups | # lines | # groups |
| 000–000 | 1639 | 207 | 53 | 544 | 84 | 888 | 110 |
| 010–000 | 1903 | 586 | 106 | 661 | 121 | 656 | 113 |
| 100–000 | 1326 | 477 | 92 | 460 | 92 | 389 | 79 |
| 001–000 | 1626 | 557 | 98 | 527 | 102 | 542 | 102 |
| 011–000 | 1221 | 378 | 71 | 408 | 80 | 435 | 82 |

This implies the orders in picking up groups would affect the categorization results. In order to fully exploit the pair identity and the smooth variation rules, it is desirable to categorize lines as much as possible into those groups where the two rules hold firmer at lower $j$ values. Therefore, the picking priority can be determined by the $j_{bd}$ values associated with the initial and final states of the groups. In general, the priority for groups in the A–A case is higher than that in the A–C and C–C cases. Meanwhile, the priority for groups with small $\Delta'$ and $\Delta''$ is higher than that with large $\Delta'$ and $\Delta''$. In practice, one needs to balance these two considerations and to set the best choice.

According to the HITRAN 2008 database, the five strongest $H_2O$ bands are the 000–000, 010–000, 100–000, 001–000, and 011–000 bands. In Table 3, we present numbers of their lines and numbers of their groups in the P, Q, and R branches for these bands. For example, for the pure rotational band there are 207, 544, and 888 lines in the P, Q, and R branches and these lines are categorized into 53, 84, and 110 groups, respectively. The largest group in this band contains 40 lines and is characterized by $j'_{1,j'} \leftarrow j''_{0,j''}$ and $j'_{0,j'} \leftarrow j''_{1,j''}$ in the R branch. Besides many large groups, there are a lot of groups containing less than four lines. These small groups do have many members, but most of their lines are not in the list of the HITRAN database because their intensities are below the threshold in developing HITRAN.

## 4. Screening $H_2O$ data listed in HITRAN 2008 with the two rules

It is well known that different spectroscopic parameters of $H_2O$ lines have different uncertainties. Among all the six parameters, the transition frequency has the highest accuracy and the line intensity follows. The induced shift and the temperature exponent are the poorest ones. Meanwhile, the air- and self-broadened half-widths are in between them. On the other hand, as mentioned above, the two rules hold with certain tolerances and the latter could vary among the spectroscopic parameters. However, the variations of the accuracy tolerance are not as large as that of the uncertainties associated with the parameters themselves. In comparison with their corresponding tolerances, the uncertainties of the transition frequency and the line intensity are much less and meanwhile, that of the other parameters are significantly larger. These comparisons enable one to exploit the two rules more properly and fruitfully. More specifically, by screening values of the transition frequency and the line intensity, one is able to pick up their possible errors, but not able to improve their accuracies because the latter is beyond the ability of the rules. For the other parameters, one is able not only to identify their errors, but also to improve their accuracies with a smoothing procedure as well.

### 4.1. Screening transition frequencies

Based on the categorizations carried out in the important bands, we have screened all their spectroscopic parameters listed in HITRAN 2008 within the individual groups. First of all, the transition frequencies of lines in many bands provided in HITRAN 2008 follow the rules perfectly. This is well expected because values of the transition frequency are solely determined by the energy levels of $H_2O$ states associated and these energy levels are well known with high accuracies. However, abnormal

behaviors do occasionally occur in the screening practices. We note that uncertainties of the energy levels of states would not cause these events because due to their accuracy tolerances, the two rules are not sensitive enough to detect such small uncertainties. In fact, by tracing origins of these strange behaviors, we have found either wrong assignments of quantum numbers for the associated $H_2O$ states or the accidental resonances are responsible for these events. For the former, although accuracies of the transition frequency in databases are very high, wrong assignments of lines sometime do happen in modern databases which could contain millions and millions lines. With the screening results, one is able easily to pick up mislabeled lines and to make necessary corrections. In fact, during a course in our checking the energy levels provided by Barber et al. [11], some wrong assignments of the states have been identified. For examples, for the 020 mode, a state with listed energy 5237.569824 $cm^{-1}$ has been mislabeled as $10_{6,0}$ and a state with 5505.318359 $cm^{-1}$ has been mislabeled as $11_{7,0}$. Based on our screening results, we are pretty sure that their assignments should be $10_{6,4}$ and $11_{6,5}$, respectively. At this stage, we would like to mention that under the auspices of IUPAC, the most updated energy levels and transition wavenumbers for $H_2{}^{16}O$ becomes available very recently [12]. Because their paper is published after the present study was completed, we have not had a chance to look at it. We hope that most of the wrong assignments existing in previous databases have been corrected there.

On the other hand, because the accidental resonances resulting from sound physics do occur, to identify them in advance is a necessary step in order to avoid misapplying the two rules. As mentioned in our previous paper [9], the accidental resonances such as Coriolis, Fermi, Darling-Dennison, and more complex resonances can cause breakdowns of the pair identity and patter similarity for the energy levels and the wave functions of $H_2O$ states and the latter are the foundation of the two rules. This implies that once these phenomena occur, one should not apply the two rules there because their applicability has already broken down. Thus, to identify these resonance phenomena in advance become important for correctly applying the two rules. Fortunately, without investigating detailed mechanisms responsible for these phenomena, screening results for the transition frequency could provide their traces. Therefore, among all the spectroscopic parameters of lines, it is better to screen the transition frequency first.

As an example, we present results in screening the transition frequency for four specified groups in the R branch of the 010–000, 000–000, 100–000, and 020–000 bands in Fig. 1(a)–(d). In order to plot the results obtained from the different bands in the same figures, the frequency values have been subtracted by the corresponding band centers. As shown in the figures, the line frequencies in the 010–000 and 000–000 bands perfectly follow the two rules. On the other hand, although most of the line frequencies in the 100–000 and 020–000 bands perfectly follow the rules too, noticeable violations do occur at $j''=9$ and 10 in Fig. 1(a),

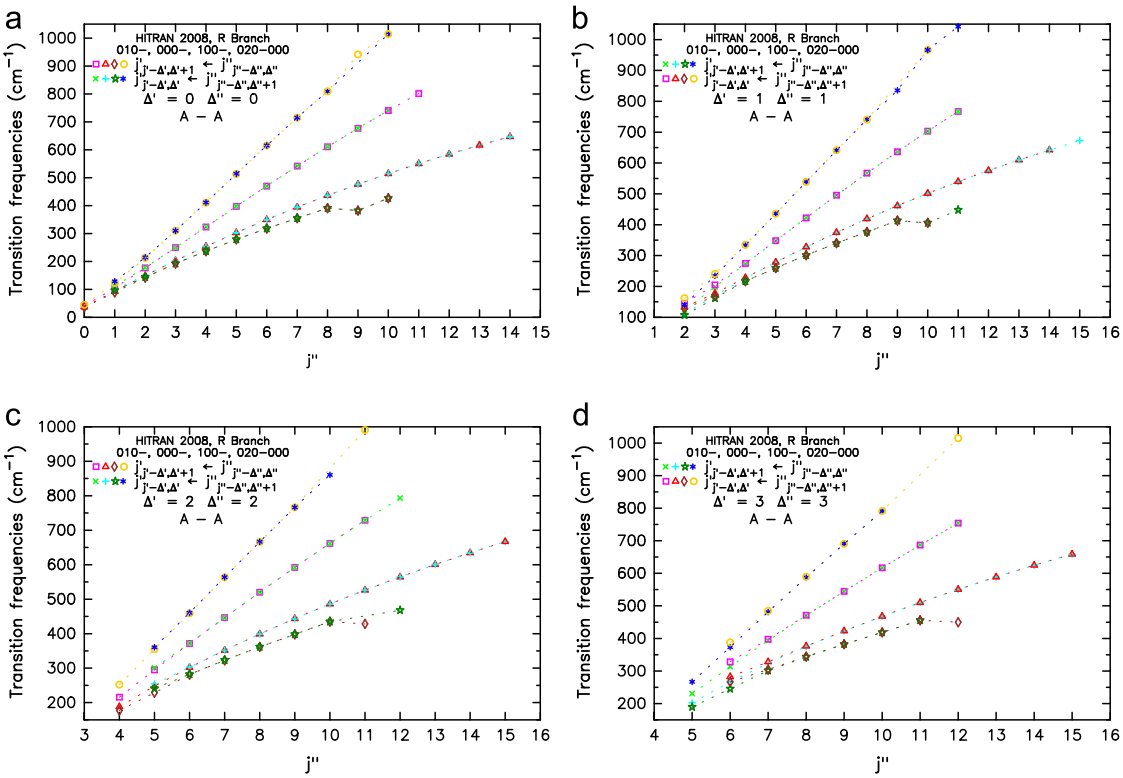


**Fig. 1.** The screening results of the transition frequency for the four groups of $j'_{j'-\Delta',\Delta'+1} \leftarrow j''_{j''-\Delta'',\Delta''}$ and $j'_{j'-\Delta',\Delta'} \leftarrow j''_{j''-\Delta'',\Delta''+1}$ in the R branch associated with the four selections of $\Delta'$ and $\Delta''$ in the 010–000, 000–000, 100–000, and 020–000 bands. The transition frequencies of these pairs subtracted by the band center in the four bands are plotted by {□, ×}, {△, +}, {◇,★}, and {○, ✻}, respectively, in Fig. 1(a)–(d).

$j''=10$ and 11 in Fig. 1(b), $j''=11$ and 12 in Fig. 1(c), and $j''=12$ in Fig. 1(d). As shown in the figures, magnitudes of these deviations could reach as large as two dozens of $cm^{-1}$ or more. Although the relative deviations are less than 1%, they are far above the 0.001 $cm^{-1}$ accuracy generally achieved in databases for the transition frequency.

Because these violations happen only for some lines in the 100–000 and 020–000 bands, it implies that they result from energy levels of their final states in the vibrational 100 and 020 modes. Then, we have checked energy levels of the states in the 100 and 020 modes provided by Barber et al. [11]. In order to show deviations more noticeably, instead of directly presenting energy levels of these states we prefer to present energy gaps between them. In Fig. 2(a) and (b), we present energy gaps between two states $j_{j-\Delta,\Delta}$ and $j_{j-\Delta-1,\Delta+1}$ with $\Delta=0,1,\ldots,4$ in the 100 and 020 modes. As shown in these figures, the energy gaps between states in the 100 and 020 modes do not vary smoothly as $j$ varies. More specifically, one significant structure occurs for each of all the curves associated with $\Delta=0,1,\ldots,4$ and their locations along the $j$ axis increase from $j=10$ to 14 with one increment successively. It is worth mentioning that the above feature is valid for both of the 100 and 020 modes, but their deviations go into opposite directions. This implies that the plots exhibit intrinsic connections of the energy levels between those specified states in the 100 and 020 modes. In fact, the structures of the energy levels simultaneously occurring at states in the 100 and 020 modes are caused by the resonance couplings between states in these two modes. It turns out that there are extremely strong centrifugal distortion resulting from bending–rotational coupling which is especially pronounced for states in 020 mode and from the resonance interactions of both anharmonic and Coriolis type between states in the 100 and 020 modes [13,14].

### 4.2. Screening line intensities

With respect to the line intensity, we prefer to present its values without the spin degeneracy factor in order to demonstrate the rules more clearly. First of all, it seems that for most of $H_2O$ lines in the five important bands, their intensities follow the pair identity and the smooth variation rules very well. As examples, we present screening results for two groups in the C–C case and in the Q branch characterized by $\Delta'=1$ and $\Delta''=0$ and by $\Delta'=2$ and $\Delta''=1$, respectively, in Fig. 3(a) and (b). The general expressions for two paired lines applicable for all these groups in the C–C case can be written as $j'_{\Delta'+\delta_0,j'-\Delta'} \leftarrow j''_{\Delta'',j''-\Delta''}$ and $j'_{\Delta'+\delta_1,j'-\Delta'} \leftarrow j''_{\Delta''+1,j''-\Delta''}$ where $\delta_0=0$ and $\delta_1=1$ for the 000–000, 010–000, and 100–000 bands and $\delta_0=1$ and $\delta_1=0$ for the 001–000 and 011–000 bands. As shown in these figures, two paired lines with their $j''$ values beyond the certain boundaries in these five bands have almost identical intensities. This demonstrates the applicability of the pair identity rule. Meanwhile, their intensities of the pairs in the same groups vary smoothly as the pairs labeled by $j''$ vary. This implies the smooth variation rule holds. As shown in the figures that the applicability of the second rule becomes valid earlier than the first rule.

However, by checking all the screening results, serious violations of the two rules have been found for many groups in these five bands. These events demonstrate that some of line intensity values in HITRAN 2008 have to be updated. In the following, we present some samples of the violations. For the groups in the R branch of the five bands, we present the screening results for the eight groups whose intensities may contain large errors in Fig.4(a)–(h). For groups in the A–A case, general expressions for the paired lines can be expressed as $j'_{j'-\Delta',\Delta'+\delta_0} \leftarrow j''_{j''-\Delta'',\Delta''}$ and $j'_{j'-\Delta',\Delta'+\delta_1} \leftarrow j''_{j''-\Delta'',\Delta''+1}$ where $\delta_0=0$ and $\delta_1=1$ for the 00–000, 010–000, and 100–000 bands and $\delta_0=1$ and $\delta_1=0$ for the 001–000 and 011–000 bands. We note that only the first and the second groups are available for all the five bands. The third group does not exist in the 000–000 band and the remaining five groups are not available for the 001–000 and 011–000 band. In cases of some bands become unavailable, their symbols and labels are absent in the plots. Generally speaking, it seems that the intensity data of the 100—000 band are poorer than other bands. By looking at Fig. 4(a) and (b), it is obvious that in contrast with other bands, variations of intensities of the 100–000 band behaves strangely, and as shown in Fig. 4(a), there is a pair of lines with $j''=13$ in this band missing and there are three partners with $j''=14$, 15, and 16 missing also.

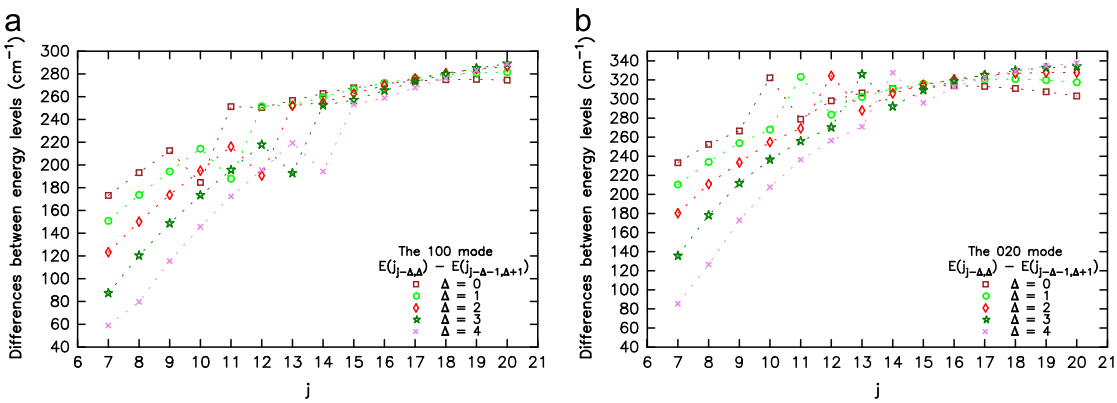


**Fig. 2.** Plots to show energy gaps between states of interest in the 100 and 020 vibrational modes. The energy gaps between two states $j_{j-\Delta,\Delta}$ and $j_{j-\Delta-1,\Delta+1}$ with $\Delta=0,1,\ldots,5$ in the 100 and 020 modes are plotted by symbols □, ○, ◇, ★, and ×, respectively, in Fig. 2(a) and (b). The same symbols are connected by dotted lines.

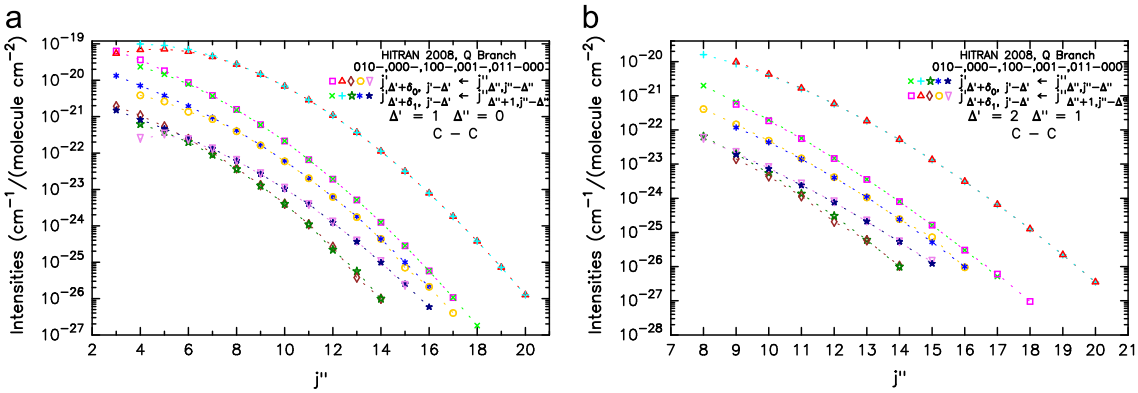


**Fig. 3.** The intensities of paired lines (in $cm^{-1}/(molecule\ cm^{-2})$) in two groups in the C–C case and in the Q branch, and with $\Delta'=1, \Delta''=0$ and $\Delta'=2, \Delta''=1$, respectively, in the five important $H_2O$ bands labeled by 000–000, 010–000, 100–000, 001–000, and 011–000 are plotted in Fig. 3(a) and (b). By excluding the spin degeneracy factor, for each of the two paired lines, their intensity values in the five bands listed in HITRAN 2008 are represented by five different symbols. In addition, the same symbols are connected by their corresponding lines.

A usual cut-off for the lines intensity cannot be used to explain all these missing lines from the linelist.

Besides, there are errors for other bands. As examples, for the 000–000 band there are large errors happening at the lines of $19_{5,15}\leftarrow18_{2,16}$, $20_{5,16}\leftarrow19_{2,17}$, and $20_{4,16}\leftarrow19_{3,17}$ in Fig. 4(b). For the 010–000 band, serious errors happen at the line $18_{5,14}\leftarrow17_{2,15}$ in Fig. 4(b), at the pair of $8_{7,1}\leftarrow7_{4,4}$ and $8_{7,2}\leftarrow7_{4,3}$ in Fig. 4(f), at $9_{7,2}\leftarrow8_{4,5}$ in Fig. 4(g), and also at $10_{7,3}\leftarrow9_{4,6}$ in Fig. 4(h).

Similarly, we present screening results for the specified eight groups in the Q and P branches in Fig. 5(a)–(h) and in Fig. 6(a)–(h). As shown in these figures, line intensities in some bands contain large errors. For shortening the manuscript, we do not explicitly indicate which of the lines whose intensities violate the two rules.

As mentioned previously, for groups in the A–A case, there are no overlaps between those available in the 000–000, 010–000, and 100–000 bands and that available in the 001–000 and 011–000 bands. In Figs. 4–6, we only present the A–A groups available in the first three bands. For the A–A groups available in the 001–000 and 011–000 bands, we present some samples in Figs. 7(a)–(h) and 8(a)–(h). As shown in these figures, serious violations of the two rules do happen in the line intensities of these two bands.

### 4.3. Screening air-broadened and self-broadened half-widths

Next, we consider the air-broadened half-width and present some screening results to show many mistakes existing in HITRAN 2008. In Fig. 9(a)–(d), we present the listed half-width values for four groups of the R branch in the 001–000 band. As shown in the figure, some of the listed values follow the pair identity and the smooth variation rules, although there are a lot of violations happening in all these groups. Violations of the two rules occur at the lines with $j''=16$ and 17 in Fig. 9(a), at lines with $j''=12$, 14, and 15 in Fig. 9(b), at lines with $j''=6$, 7, and 8 in Fig. 9(c), and at lines with $j''=11$, 13, and 14 in Fig. 9(d). In general, the higher the $j''$, the more likely the violation happens. This is in contradiction to the fact that the higher the $j$, the firmer the rules hold. In order to find origins of these violations, we have checked the source code of the air-broadened half-width in HITRAN 2008 and have found that the data in Fig. 9(a)–(d) come from several different sources. Some are measured values provided by different labs [15–17] and others are derived from a semi-empirical method [8]. We have found that in many cases, the breakdowns of the pair identity rule for paired lines are caused by adopting values from two different sources. Meanwhile, some of values derived from the semi-empirical method not only violate the smooth variation rule, but also violate the pair identity rule. This implies that the applicability of this method is questionable. It is worth mentioning that the above analysis provides some useful advice about how to select data from multiple sources. It is better to adopt data from the same source for lines within the same groups, especially for two paired lines. Otherwise, one needs to consider how to balance values obtained from different sources.

Similarly, we consider the self-broadened half-widths listed in HITRAN 2008. In Fig. 10(a)–(d), we present the listed values for the four groups of the Q branch in the 011–000 band. As shown in the figures, violations of the pair identity and the smooth variation rules become more severe. For some pairs, relative differences of their values could be beyond 100%. Because the large gaps of half-width data between two paired lines somehow reflect uncertainties of these values, the plots shown in Fig. 10 clearly demonstrate the poor quality of the data in the 011–000 band. By checking the source code, we have found that the majority of these values are from two different measurements [16,17] and others are from theoretical analyses of measurements. Again, to adopt values from two different sources for paired lines is a main reason responsible for the violation of the pair identity rule.

### 4.4. Screening pressure-induced shifts

Then, we present screening results of the induced shift for eight groups in the pure rotational band Fig. 11(a)–(h) where most of the shift data come from the same theoretical

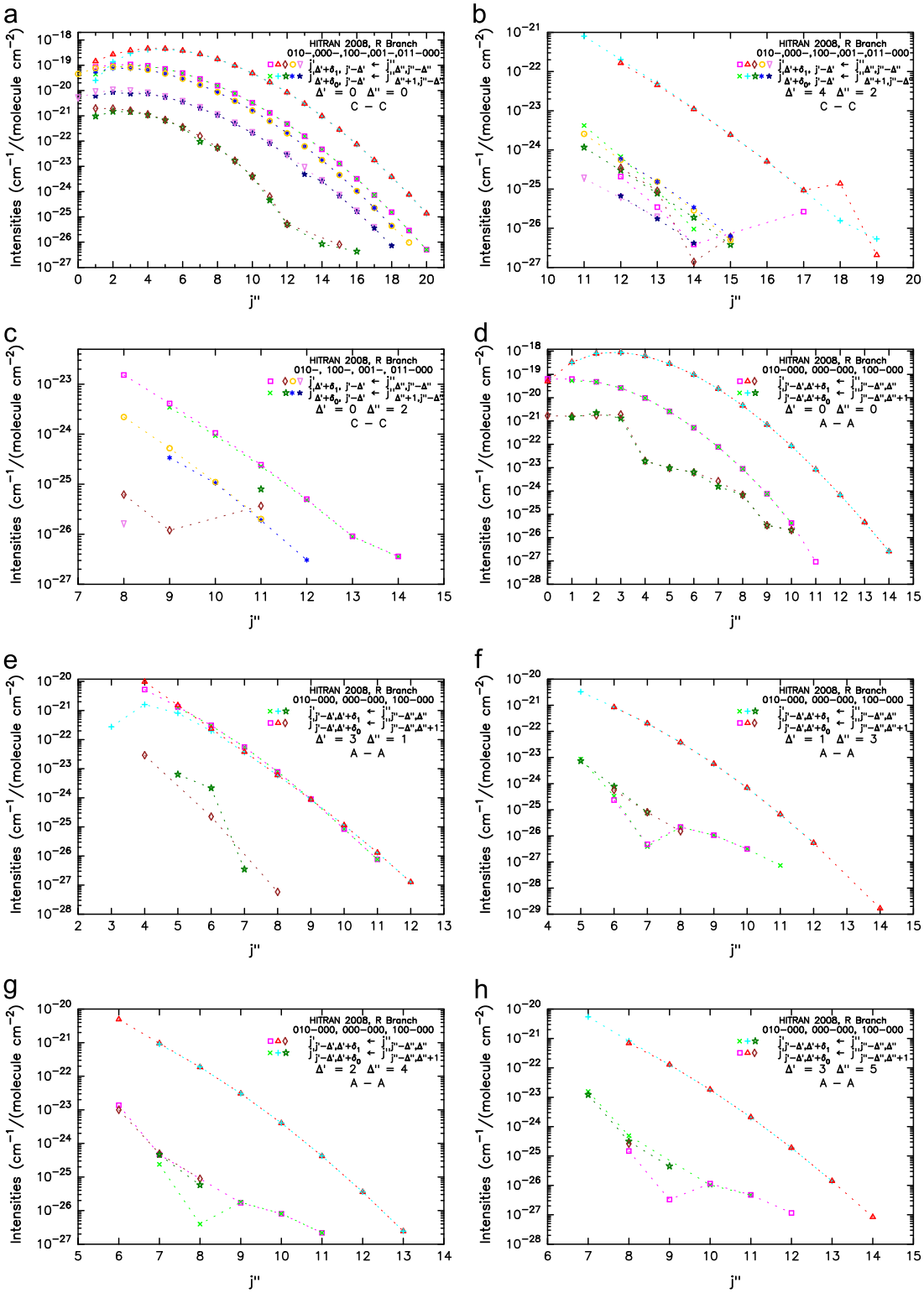


**Fig. 4.** The same as Fig. 3 except for the eight specified groups of lines whose intensities are plotted Fig. 4(a)–(h). In addition, these eight groups are not always available in all the five bands. In case the groups of interest become unavailable in some bands, the corresponding symbols and labels are absent in the plotting.

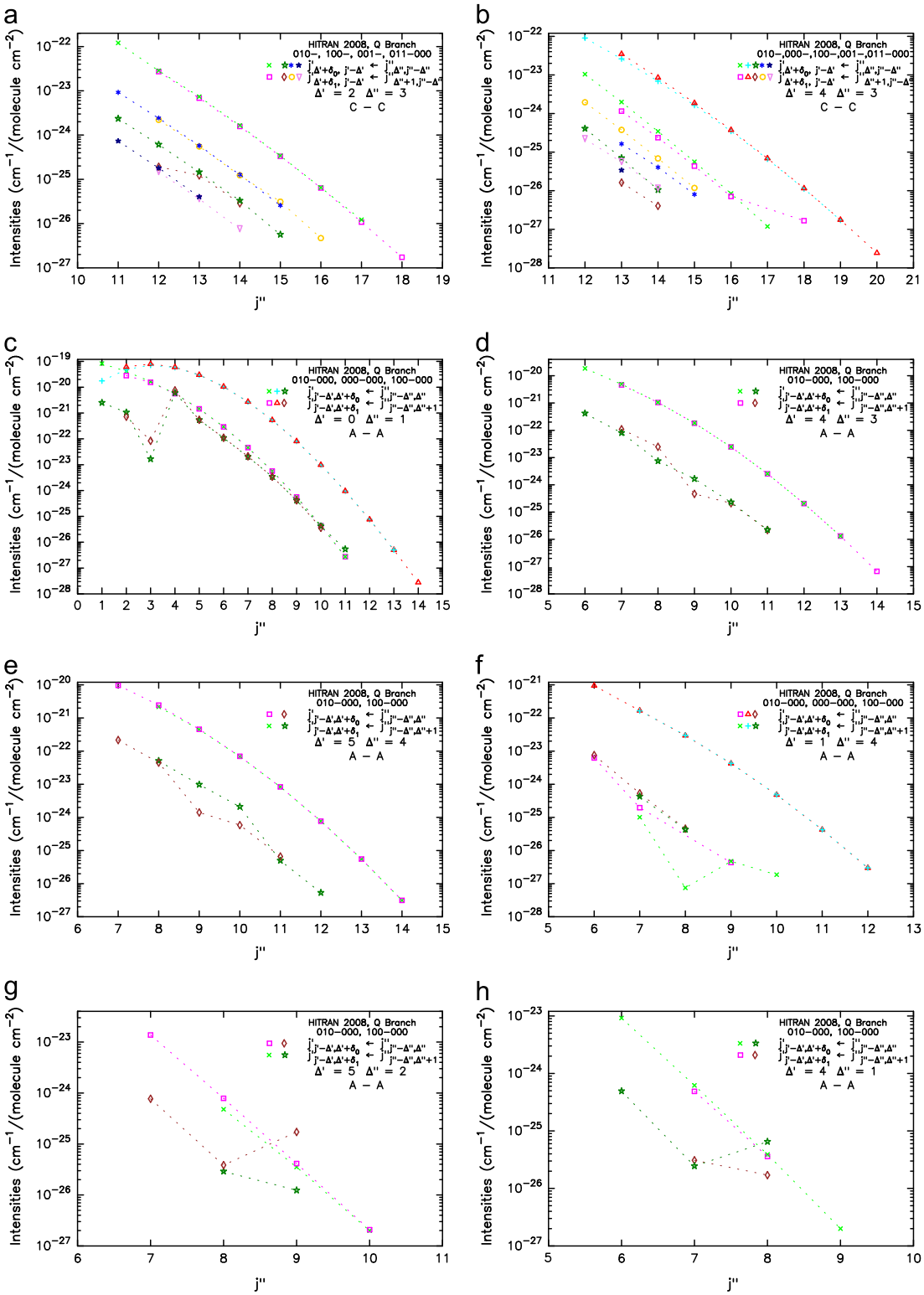


**Fig. 5.** The same as Fig. 4 except for eight specified groups in the Q branch.

calculations [7]. As shown in the figures, in general the data follow the pair identity rule. However, there are serious violations of the smooth variation rule. Because the adoption of multiple sources does not play an important role, it is the theoretical calculation itself that results in the violations. As a comparison, we present our calculated values based on the

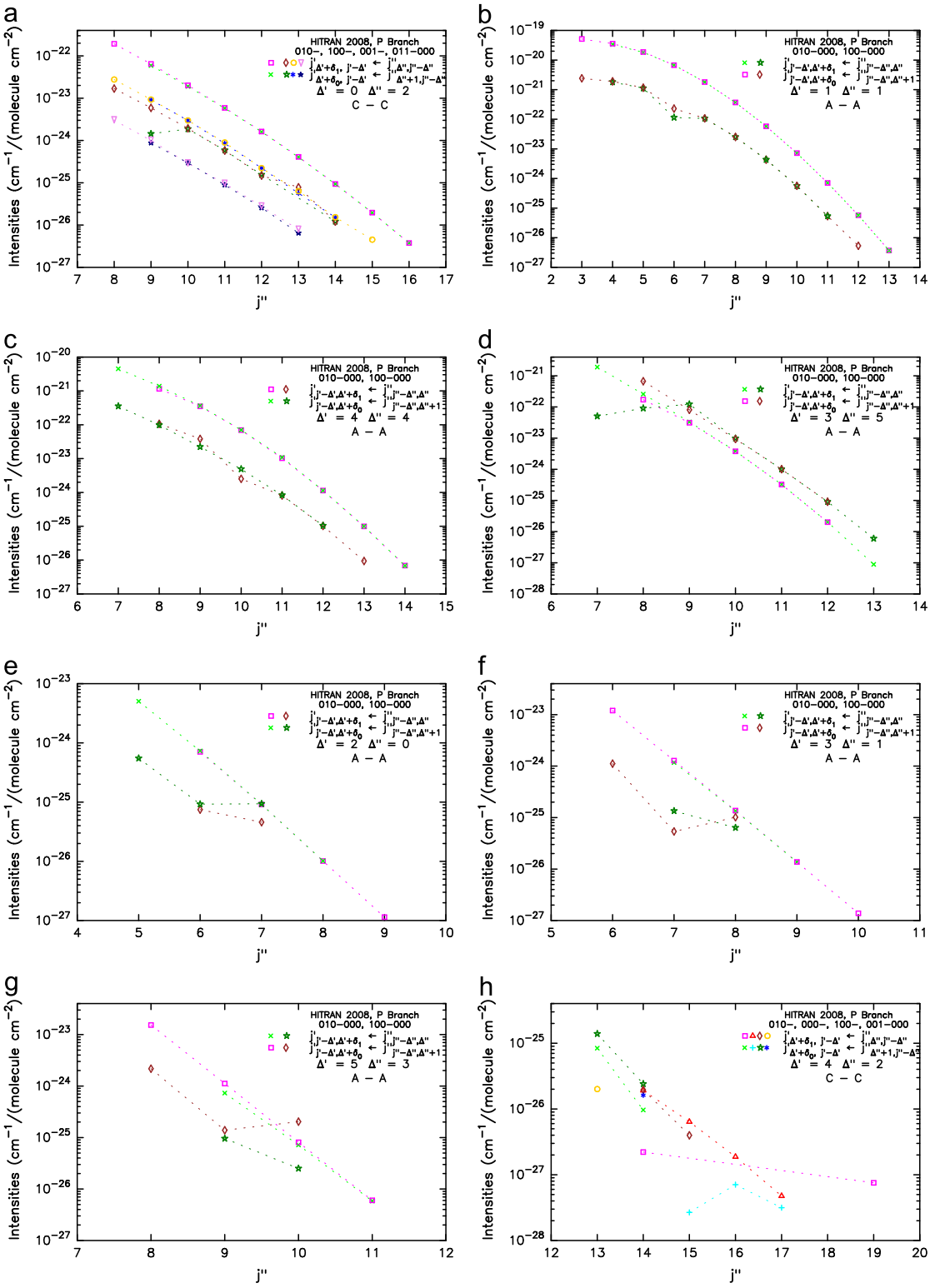


**Fig. 6.** The same as Fig. 4 except for eight specified groups in the P branch.

same potential model used in Ref. [7] in Fig. 11(a)–(h). As shown in the figures, our calculated values differ significantly from those listed in HITRAN 2008 and most importantly, our results follow the two rules very well. Thus, the severe violations of the two rules demonstrated in Fig. 11 definitely mean large mistakes exist in these shift data.

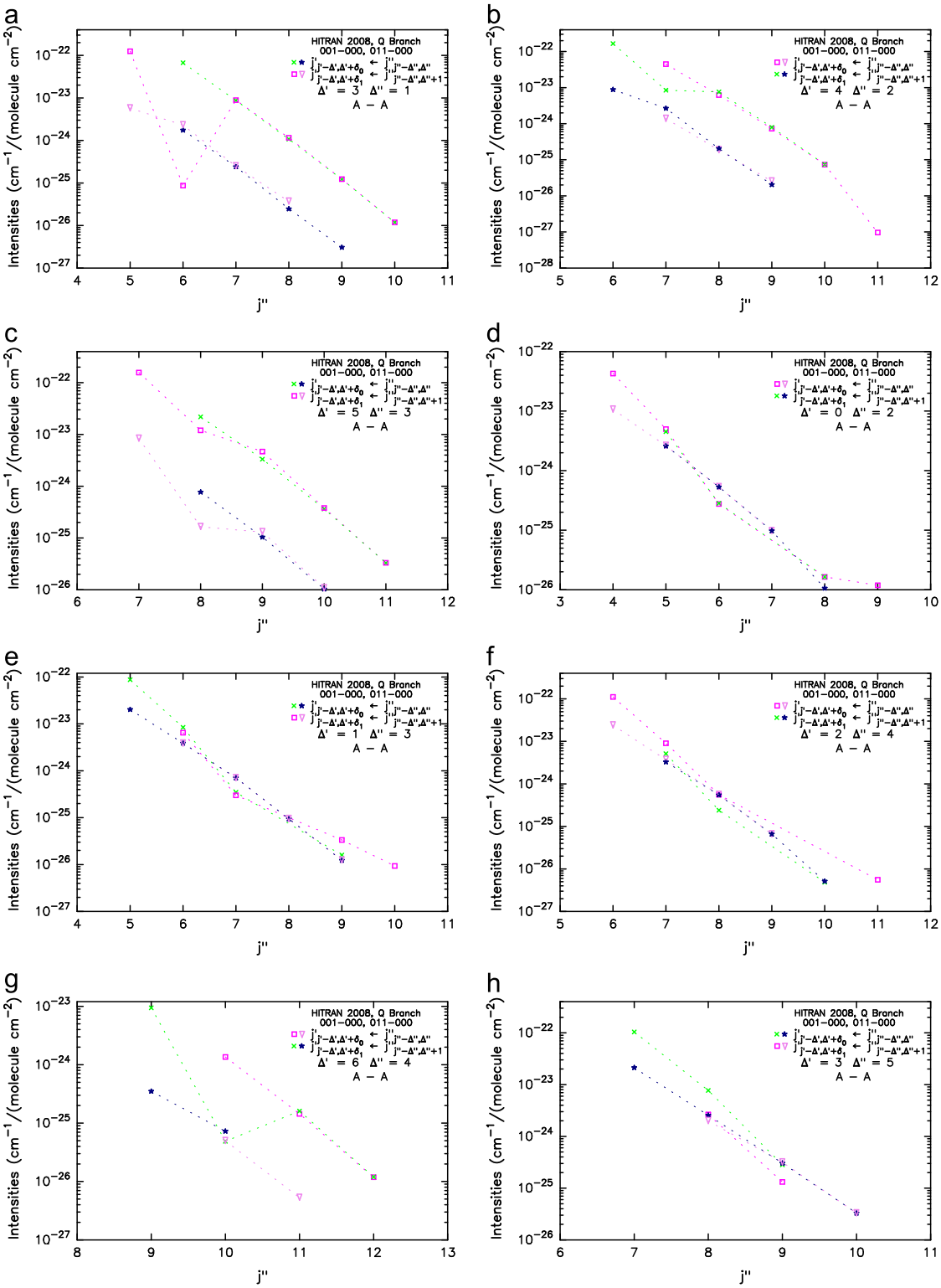


**Fig. 7.** The same as Fig. 4 except for eight specified groups in the Q branch of the 001–000 and 011–000 bands.

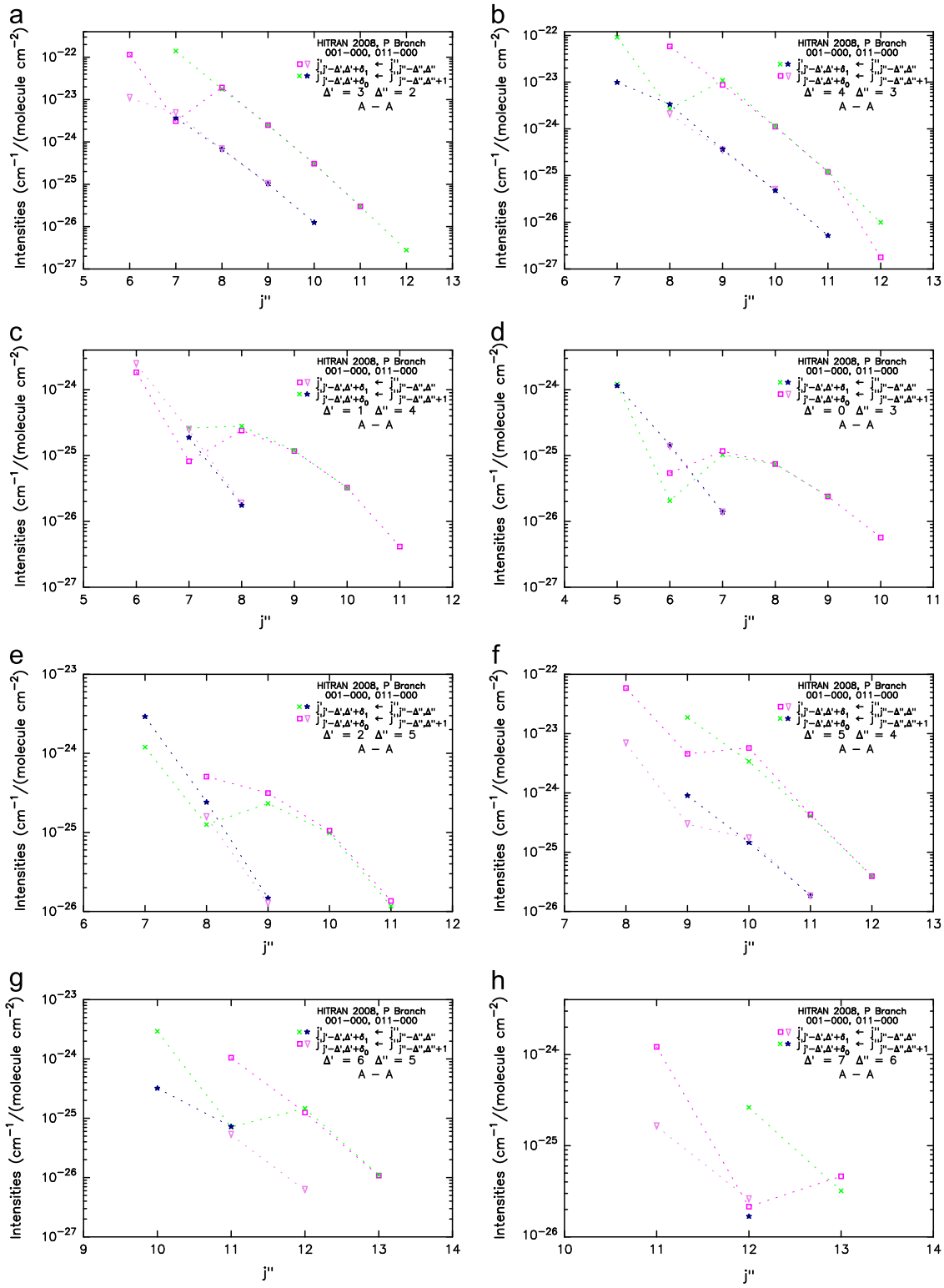


**Fig. 8.** The same as Fig. 4 except for eight specified groups in the P branch of the 001–000 and 011–000 bands.

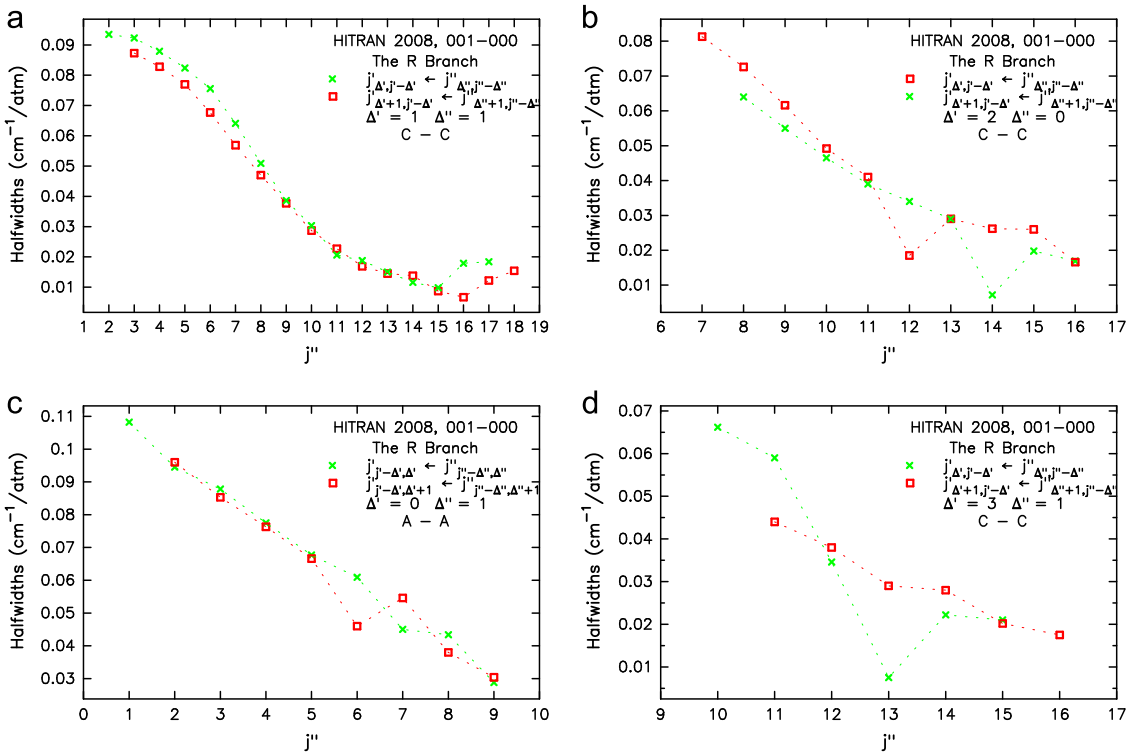


**Fig. 9.** Air-broadened half-widths (in units of $cm^{-1}$/atm) associated with four specified groups in the R branch of the 001–000 band are plotted in Fig. 9 (a)–(d). Their values listed in HITRAN 2008 are represented by □ and ×, respectively, and the same symbols are connected by their corresponding lines.

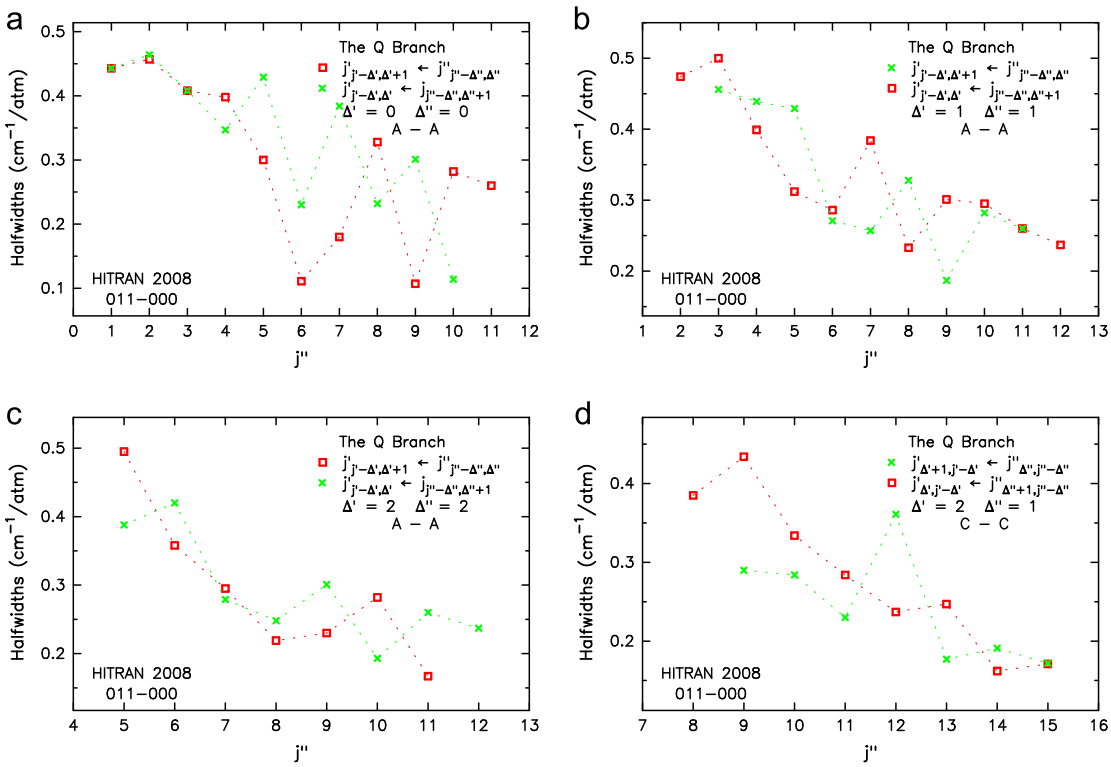


**Fig. 10.** The same as Fig. 9 except for self-broadened half-widths associated with four specified groups in the Q branch of the 011–000 band.

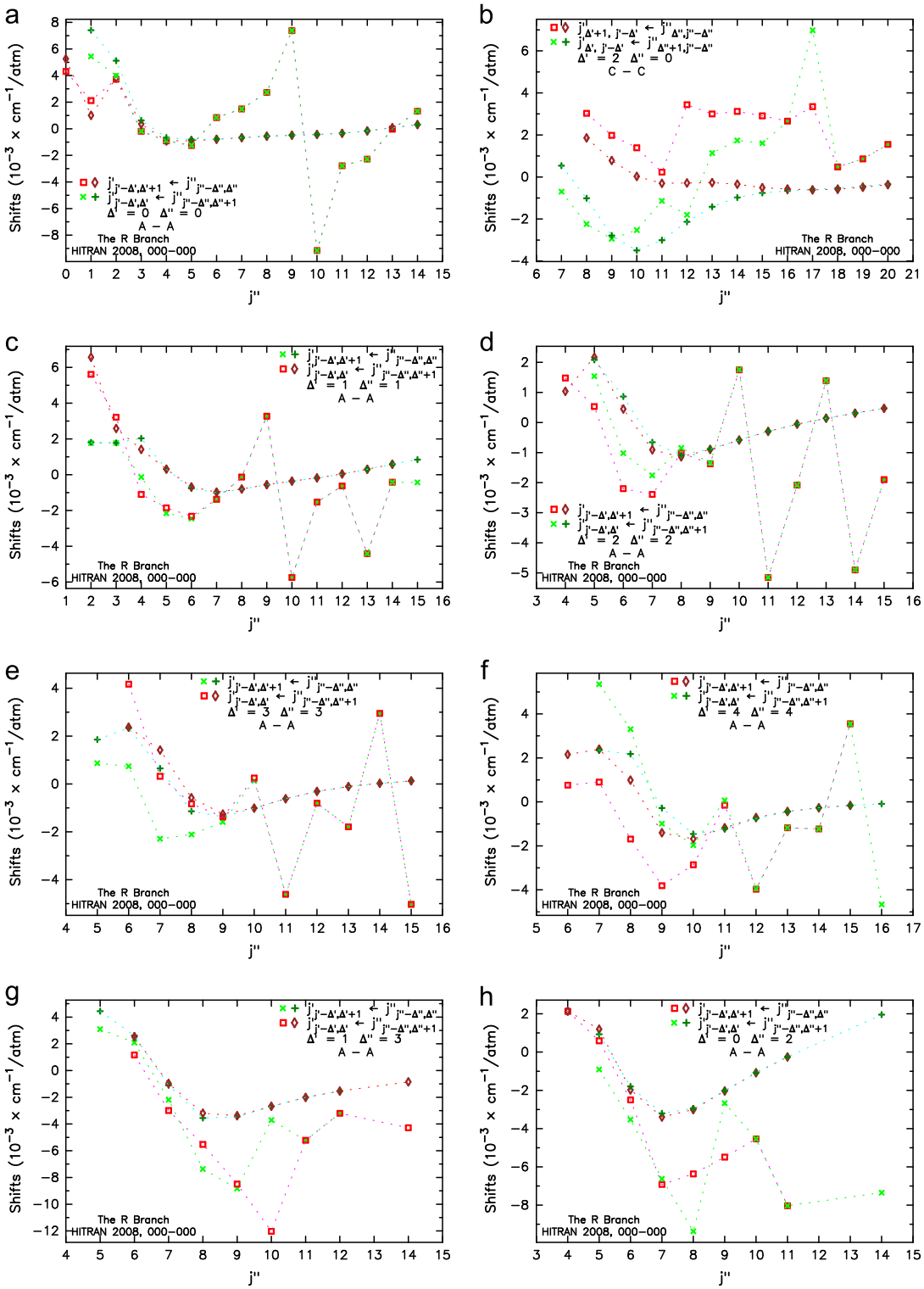


**Fig. 11.** The same as Fig. 9 except for pressure induced shifts (in units of $10^{-3} \times \mathrm{cm}^{-1}$/atm) for eight specified groups in the R branch of the pure rotational band. The values listed in HITRAN 2008 are represented by □ and ×. Meanwhile, our calculated results are given by ◊ and +.

We do not present screening results for the temperature exponent here. In general, the results follow the two rules. However, there are a lot of violations, especially the smooth variation rule. This implies that it is necessary to update these values listed in HITRAN 2008.

## 5. Verification of $H_2O$ linelists

After the screening results for all the lines in the important bands are available, not only we can pick up mistakes, but also we can identify missing lines by going through all individual figures to look at variations of the line intensity within the groups. As examples, we consider the screening results for the pure rotational band here.

### 5.1. Picking up questionable intensity data

We present four samples to demonstrate how to pick up questionable intensity data in Fig. 12(a)–(d). As shown in Fig. 12(a), there are violations of the pair identity rule occurring in two pairs with $j''=18$ and 19 in the first specified group. By checking their source code, all of them come from measured values [16]. But, the absence of the line $21_{4,18}\leftarrow 20_{1,19}$ which is the partner of the line $21_{3,18}\leftarrow 20_{2,19}$ is justified because after including the spin degeneracy factor the former is weaker than the latter by a factor of 3 and it is below the threshold. With respect to the second group shown in Fig. 12(b), there are serious violations for both the pair identity and the smooth variation rules occurring in the last two pairs with $j''=18$ and 19. The mistake of the line $19_{5,15}\leftarrow 18_{2,16}$ is strikingly large. Again, by checking their source code, all of them come from the same source [16].

With respect to the third group, it seems that the intensity value of $19_{6,14}\leftarrow 18_{3,15}$ (i.e., 1.342E−26 in $cm^{-1}/(molecule\ cm^{-2})$ ) is questionable. As shown in Fig. 12(c), this value not only violates the pair identity rule, but also violates the smooth variation rule. For the fourth group presented in Fig. 12(d), it is obvious that both the intensities of $19_{6,14}\leftarrow 18_{1,17}$ and $19_{5,14}\leftarrow 18_{2,17}$ are not correct.

### 5.2. Identifying missing lines

Then, by presenting four samples in Fig. 13(a)–(d), we show how to identify missing lines. For the first group shown in Fig. 13(a), predicted intensities of the paired lines with $j''=21$, 22, and 23 are plotted by symbol ⋆. There are five lines whose intensities are above the threshold. We note that $22_{1,22}\leftarrow 21_{0,21}$ is the strongest one and its intensity is two orders above the threshold. With respect to the second group shown in Fig. 13(b) where all the listed intensities are measured values [16], the situation becomes more complicated. First of all, one can conclude without any doubt that the intensity value 9.682E−26 (in $cm^{-1}/(molecule\ cm^{-2})$) of the line $19_{5,15}\leftarrow 18_{0,18}$ is completely wrong. Secondly, as indicated

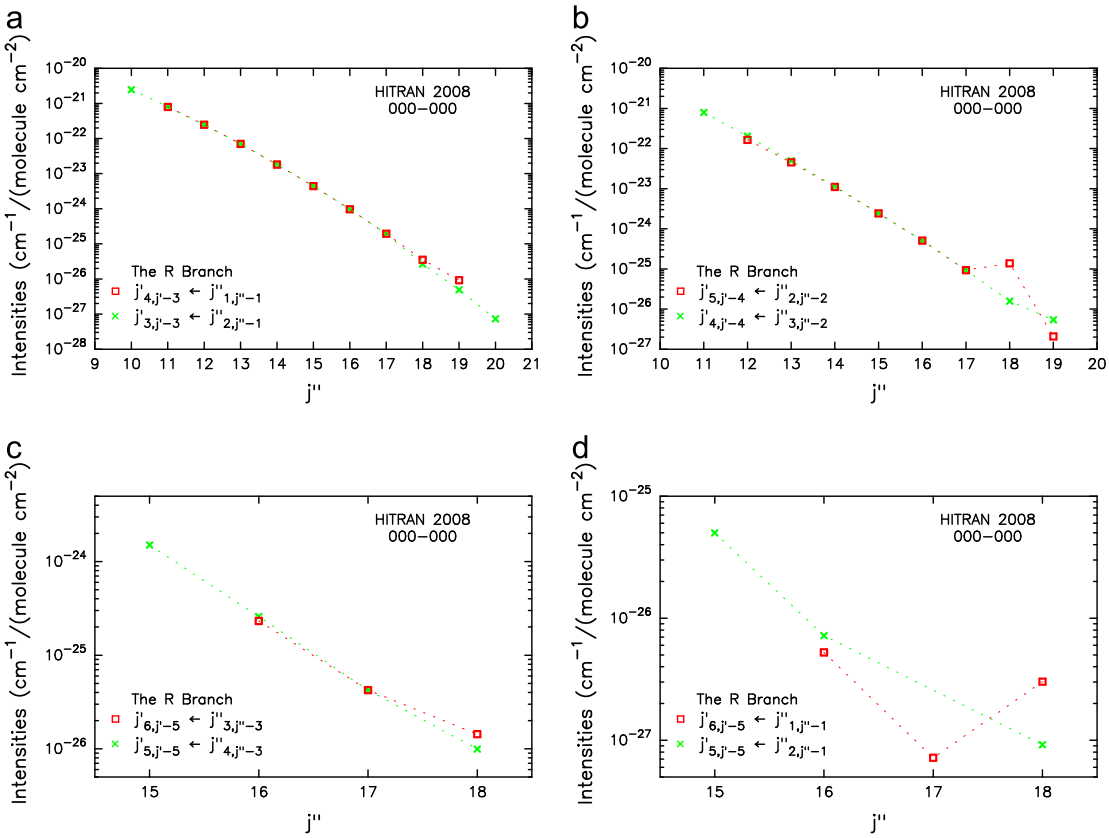


**Fig. 12.** The same as Fig. 3 except for four specified groups in the pure rotational band whose intensities values are questionable.

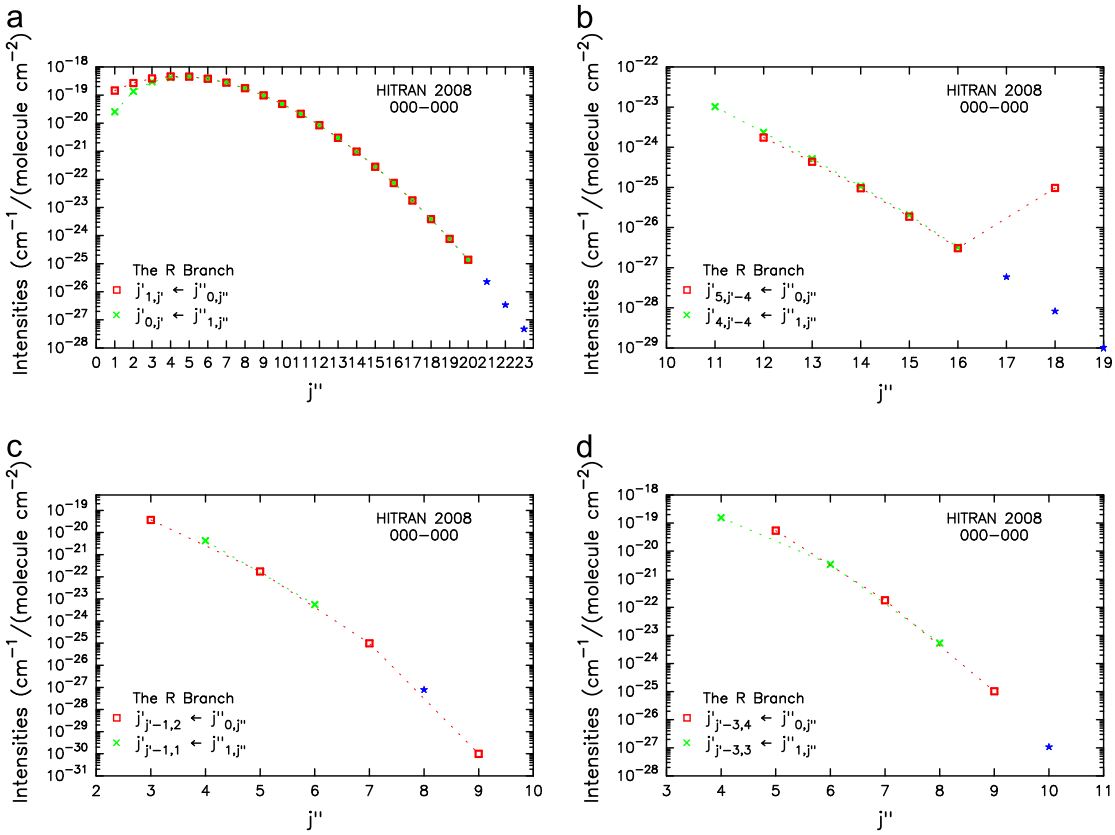


**Fig. 13.** The same as Fig. 3 except for four specified groups in the pure rotational band whose missing lines are explicitly indicated.

by a gap between the pair with $j''=16$ and the one line with $j''=18$, it is obvious that there are two paired lines missing and they are lines of $18_{5,14}\leftarrow17_{0,17}$ (at 1451.500 cm$^{-1}$) and $18_{4,14}\leftarrow17_{1,17}$ (at 1445.806 cm$^{-1}$). By using an extrapolation method based on those values with $j''\leq16$, we can estimate intensities of the two pairs with $j''=17$ and 18. We present the estimated results by symbol ⋆ in the figure. It turns out that after including spin degeneracy factor, the predicted intensities for the paired lines with $j''=17$ are 1.757E−27 and 5.858E−28, respectively. Because both of them are below the threshold, to exclude these two lines in the linelist is justified. Furthermore, the predicted value for the line $19_{5,15}\leftarrow18_{0,18}$ is 8.217E−29 which is even weaker. The difference between the value listed in HITRAN 2008 and the predicted one is as large as three orders. As a result, the listed value for $19_{5,15}\leftarrow18_{0,18}$ is completely wrong and one must remove this line from the linelist.

In the last two groups, the line missing also occurs. In contrast with the first two groups belonging to the C–C case, these two groups are in the A–C case. As explained previously, due the selection rule there is only one transition allowed for each of the pairs. With respect to the group shown in Fig. 13(c), there is one line of $9_{8,1}\leftarrow8_{1,8}$ missing as indicated by symbol ⋆. It is interesting to note that this line was listed in early HITRAN versions at 1265.64160 cm$^{-1}$ with the intensity 1.46E−26 and it somehow had been removed from the recent HITRAN versions. Meanwhile, the line of $10_{9,2}\leftarrow9_{0,9}$ should be removed because it is too weak. Finally, with the extrapolation, we can find there is one line missing in Fig. 13(d). We list all these missing lines in Table 4.

**Table 4**
Missing lines identified in Fig. 13.

| Groups | Assignments of missing lines | Frequencies (cm$^{-1}$) | Estimated intensities (cm$^{-1}$/molecule cm$^{-2}$) |
|---|---|---|---|
| Fig. 13(a) | $22_{1,22}\leftarrow21_{0,21}$ | 407.747 | 6.774E−26 |
| | $22_{0,22}\leftarrow21_{1,21}$ | 407.747 | 2.258E−26 |
| | $23_{1,23}\leftarrow22_{0,22}$ | 424.865 | 3.392E−27 |
| | $23_{0,23}\leftarrow22_{1,22}$ | 424.875 | 1.018E−26 |
| | $24_{1,24}\leftarrow23_{0,23}$ | 441.879 | 1.398E−27 |
| Fig. 13(c) | $9_{8,1}\leftarrow8_{1,8}$ | 1265.642 | 1.975E−27 |
| Fig. 13(d) | $11_{8,3}\leftarrow10_{1,10}$ | 1407.715 | 3.210E−27 |

### 5.3. Providing supplement of the missing line list

At this stage, we would like to emphasize that to identify missing lines and to provide their estimated spectroscopic parameters are two tasks with different achievements of the certainty. One can complete the first task with a pretty high certainty. Because the groups of interest are well characterized from the categorization procedure, we are pretty sure that correctly assigning the

quantum numbers for the missing lines can be well achieved. On the other hand, whether one is able to provide reliable spectroscopic parameters for the missing lines depends on several factors including the parameters themselves. For the transition frequency, one is able to provide very accurate values for the missing lines because these values can be easily calculated from energy levels associated with their initial and final states and the latter's high accurate values are available in literature. As a result, one is able to provide basic missing linelists with high certainty that consist of the line assignments and positions.

For other parameters including the line intensity, to provide their reliable estimated values is a big challenge. In these cases, one has to rely on interpolation or extrapolation methods based on original data available in the same groups. How accurate the estimated values are depends on several factors: how good and how many the original data themselves are and how suitable the extrapolation methods are. It is well known that the extrapolation methods could cause large uncertainties and different methods could yield quite different results. Therefore, in general, estimated values for the other spectroscopic parameters are less reliable. At present, although one can provide compete supplements for missing lines consisting of all their parameter values, it is better to consider them as tentative because they contain large uncertainties.

After presenting a demonstration to show how to find missing lines in Fig. 13(a)–(d), we present missing lines in the pure rotational band identified from most of the groups. By going through all the groups in the pure rotational band which contain at least four lines and carrying out the interpolation or the extrapolation procedure for each of them, we have found 65 lines which are not in the linelist and whose intensities are above the threshold. The assignments of these missing 65 lines together with their frequencies and estimated intensities are provided in Appendix A. For smaller groups containing two or three lines, to derive missing lines can also been done that results in adding another 16 missing ones. We note that the total 81 missing lines are not complete because there are 36 groups which contain only one line.

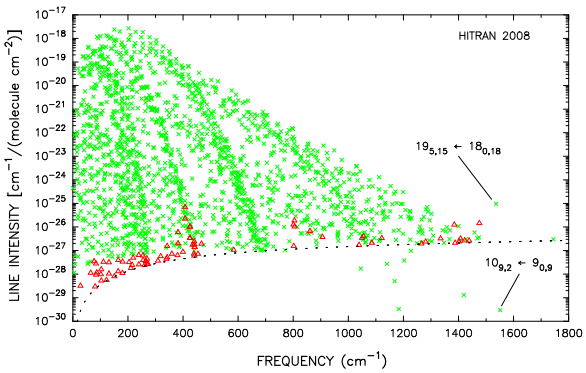


**Fig. 14.** The distribution of 81 missing $H_2O$ lines in the pure rotational band over the transition positions. Their predicted intensities (in cm$^{-1}$/(molecule cm$^{-2}$)) are plotted by △. The original 1639 $H_2O$ lines listed in HITRAN 2008 are also plotted by ×. The intensity threshold is plotted by a dotted line.

Besides, we have skipped some groups whose listed intensity values are too poor such that one cannot perform a meaningful interpolation or extrapolation at all. As a result, the total number of missing lines could be more. In Fig. 14, we represent a spectrum of these missing 81 lines together with the spectrum of the 1639 lines listed in HITRAN 2008. As shown in the figure, in comparison with the original lines, the missing ones are relatively weaker. However, some missing lines located at around 400 cm$^{-1}$ are pretty strong. In addition, after removing the wrong line of $19_{5,15} \leftarrow 18_{0,18}$ explicitly indicated in Fig. 14, the missing lines become comparable with the original ones in a region beyond 1400 cm$^{-1}$. Therefore, to include these missing lines is necessary in updating the database. On the other hand, as shown in the figure, there are 8 listed lines, including $10_{9,2} \leftarrow 9_{0,9}$ mentioned above and explicitly indicated here, whose intensities are below the threshold. In order to keep consistency, one needs to remove them from the linelist.

## 6. Discussions and conclusions

The molecular spectroscopic databases such as HITRAN are an essential resource in atmospheric applications. Researchers in many different disciplines have come to expect the high quality of these databases. During the past several decades, dramatic efforts have been made to improve the completeness and the accuracy of the linelists for a lot of atmospheric molecules, especially for the most important $H_2O$ molecule. However, there are still large gaps between what have been achieved and what are required by the users and it seems that one cannot meet the users' requirements within a short period of the time. As a result, simple and effective ways to improve the databases are more than welcome. The present work is one of such efforts.

Our approach is based on the following considerations. First of all, based on the black box theory, we believe that there are intrinsic correlations between the lines and their spectroscopic parameters and the ability to grasp these correlations is a key to develop the tool. Secondly, we know that the discovery of the correlations requires answering two fundamental questions: why and how the spectroscopic parameters of the $H_2O$ transitions vary with the lines of interest. The questions lead us into analyzing properties of the energy levels and wave functions of the $H_2O$ states because the $H_2O$ lines are completely described by the latter. Finally, we have realized that one must be realistic in finding rules governing the correlations. Due to the complexity of the whole transition process involved, the correlations are very difficult to be fully grasped. This implies that it is almost impossible to set a goal to find simple common rules with which the spectroscopic parameters can be well monitored for any lines of interest unless one narrows the variation ranges of these lines. To realize this fact is essential in our success. In the present study, to narrow the variation ranges of lines is achieved by categorizing the $H_2O$ lines for each of the P, Q, and R branches such that there is only one independent variable (i.e., the initial quantum number $j''$) left to distinguish lines of interest within individually defined groups.

**Table A1**
List of the 65 missing lines.

| Assignment | | Frequency ($cm^{-1}$) | Intensity (cm/molec) | Assignment | | Frequency ($cm^{-1}$) | Intensity (cm/molec) |
|---|---|---|---|---|---|---|---|
| $j'$ $k_a'$ $k_c'$ | $j''$ $k_a''k_c''$ | | | $j'$ $k_a'$ $k_c'$ | $j''$ $k_a''k_c''$ | | |
| 13 11 3 | 12 12 0 | 64.7917 | 1.127 – 28 | 13 10 3 | 14 7 8 | 319.0369 | 5.766 – 28 |
| 17 4 13 | 18 3 16 | 77.3210 | 3.504 – 28 | 12 10 2 | 13 7 7 | 339.6892 | 4.447 – 28 |
| 14 12 3 | 13 13 0 | 81.0474 | 2.956 – 29 | 21 2 19 | 21 1 20 | 342.4390 | 1.550 – 27 |
| 14 11 3 | 13 12 2 | 84.1714 | 9.368 – 29 | 21 3 19 | 21 2 20 | 342.4453 | 5.168 – 28 |
| 14 11 4 | 13 12 1 | 84.1714 | 2.810 – 28 | 11 10 2 | 12 7 5 | 359.7229 | 7.250 – 28 |
| 17 5 13 | 18 2 16 | 86.9614 | 1.168 – 28 | 10 10 0 | 11 7 5 | 380.0754 | 6.310 – 28 |
| 18 5 14 | 19 2 17 | 101.7925 | 9.891 – 29 | 21 1 20 | 21 0 21 | 381.8945 | 6.027 – 27 |
| 15 11 4 | 14 12 3 | 103.3257 | 1.593 – 28 | 21 2 20 | 21 1 21 | 381.8945 | 2.009 – 27 |
| 15 11 5 | 14 12 2 | 103.3276 | 5.309 – 29 | 22 2 21 | 22 1 22 | 399.5840 | 8.790 – 28 |
| 15 10 6 | 14 11 3 | 111.1538 | 3.311 – 28 | 22 0 22 | 21 1 21 | 407.7471 | 2.258 – 26 |
| 16 11 6 | 15 12 3 | 122.2817 | 5.613 – 29 | 22 1 22 | 21 0 21 | 407.7471 | 6.774 – 26 |
| 16 10 7 | 15 11 4 | 131.0103 | 3.150 – 28 | 23 1 23 | 22 0 22 | 424.8652 | 3.392 – 27 |
| 16 10 6 | 15 11 5 | 131.0171 | 1.050 – 28 | 23 0 23 | 22 1 22 | 424.8750 | 1.018 – 26 |
| 17 10 7 | 16 11 6 | 150.7788 | 8.043 – 29 | 22 1 21 | 21 2 20 | 425.4272 | 3.410 – 27 |
| 17 9 9 | 16 10 6 | 164.6318 | 1.312 – 28 | 22 2 21 | 21 1 20 | 425.4365 | 1.023 – 26 |
| 17 9 8 | 16 10 7 | 164.9200 | 3.936 – 28 | 22 2 20 | 21 3 19 | 441.2188 | 6.109 – 28 |
| 18 3 15 | 19 2 18 | 180.0361 | 1.119 – 28 | 22 3 20 | 21 2 19 | 441.2280 | 1.833 – 27 |
| 18 4 15 | 19 1 18 | 180.6399 | 3.357 – 28 | 24 1 24 | 23 0 23 | 441.8794 | 1.398 – 27 |
| 18 8 11 | 17 9 8 | 204.2324 | 3.204 – 28 | 23 1 22 | 22 2 21 | 442.4229 | 1.416 – 27 |
| 18 9 9 | 18 8 10 | 215.2583 | 2.438 – 28 | 21 4 17 | 20 5 16 | 454.0835 | 7.236 – 28 |
| 19 6 14 | 19 5 15 | 219.0693 | 4.913 – 28 | 15 4 11 | 16 1 16 | 583.6548 | 1.064 – 27 |
| 17 10 8 | 17 9 9 | 239.3965 | 6.433 – 28 | 8 8 0 | 8 1 7 | 906.1527 | 3.736 – 27 |
| 18 10 9 | 18 9 10 | 239.6421 | 2.577 – 28 | 12 8 4 | 12 1 11 | 1038.9124 | 1.720 – 27 |
| 17 11 6 | 17 10 7 | 254.6333 | 6.141 – 28 | 12 7 5 | 12 0 12 | 1055.2601 | 4.056 – 27 |
| 18 7 11 | 17 8 10 | 255.2036 | 3.963 – 28 | 13 8 5 | 13 1 12 | 1085.5511 | 2.126 – 27 |
| 20 4 16 | 20 3 17 | 260.3086 | 2.425 – 28 | 13 7 6 | 13 0 13 | 1121.2712 | 3.354 – 27 |
| 16 12 4 | 16 11 5 | 264.3384 | 3.550 – 28 | 9 8 1 | 8 1 8 | 1265.6425 | 1.975 – 27 |
| 18 2 16 | 19 1 19 | 265.4316 | 3.523 – 28 | 12 10 2 | 11 5 7 | 1280.9796 | 2.367 – 27 |
| 15 13 3 | 15 12 4 | 269.2568 | 4.463 – 28 | 13 9 5 | 12 2 10 | 1387.5729 | 2.112 – 27 |
| 14 14 0 | 14 13 1 | 269.3799 | 2.768 – 28 | 11 8 3 | 10 1 10 | 1407.7153 | 3.210 – 27 |
| 16 13 4 | 16 12 5 | 272.2983 | 2.644 – 28 | 14 8 7 | 13 1 12 | 1422.5770 | 2.666 – 27 |
| 15 14 1 | 15 13 2 | 273.4434 | 2.674 – 28 | 14 9 6 | 13 1 11 | 1438.5232 | 2.636 – 27 |
| 21 3 18 | 21 2 19 | 308.7822 | 3.708 – 28 | | | | |

For the $H_2O$ molecule, the categorization procedure is well guided by the knowledge developed in analyzing the properties of energy levels and wave functions of the $H_2O$ states such that the energy levels and wave functions associated with their initial and final states of the $H_2O$ lines belonging to the same categorized groups have similar properties. Then, based on the black box theory, the outputs (i.e., their spectroscopic parameters) must share the similarity with the inputs (i.e., the energy levels and wave functions associated). This implies that within the same groups, variations of the spectroscopic parameters can be well monitored by the variable $j''$ and the latter is a foundation to establish the pair identity and smooth variation rules which are the tools we are looking for. Therefore, these two rules bear two characters: they are natural, but they are local. Local here means the rules work for each of individual groups and are valid for its members above certain boundaries.

The two rules provide a helpful tool in improving databases such as HITRAN. By screening the spectroscopic parameters with them, one is able to identify possible errors and to improve the accuracies. Meanwhile, one can find missing lines in the linelist whose intensities are above the threshold. As demonstrated in Sections 4 and 5, one can make significant improvement of the HITRAN database. In comparison with other solutions, this method is doable now, and at a less cost. In addition, the methods can serve as an effective way for experimenters to identify possible errors in their measurements. Violations of the pair identity rule happening at paired lines with high $j$ values definitely mean warning signs. The outliers should be understood as red flags. In comparison with other data belonging to the same groups, once the outliers have been identified, they need to be double checked.

The present study is carried out in the most important five bands of $H_2O$, but one can extend the study to other $H_2O$ bands. In addition, the basic idea to exploit properties of the energy levels and wave functions of molecular states and to link these properties to spectroscopic parameters of their lines is general. As a result, one can consider its applications for other molecules which are important for atmospheric applications.

## Acknowledgments

Two of the authors (Q. Ma and R.H. Tipping) acknowledge financial support from NSF under Grant 1228861. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

## Appendix A

In Table A1, we provide the assignments of the missing 65 lines together with their frequencies and estimated intensities. For the intensity values, 1.127 − 28 for the line $13_{11,3} \leftarrow 12_{12,0}$ means 1.127E − 28 and so on. Readers must keep in mind that the estimated intensities contain large uncertainties.

## References


[1] Rothman LS, Jacquemart D, Barbe A, Benner DC, Birk M, Brown LR, et al. The HITRAN 2004 molecular spectroscopic database. J Quant Spectrosc Radiat Transfer 2005;96:139–204.

[2] Gordon IE, Rothman LS, Gamache RR, Jacquemart D, Boone C, Bernath PF, et al. Current updates of the water-vapor line list in HITRAN: a newdiet for air-broadened half-widths. J Quant Spectrosc Radiat Transfer 2007;108:389–402.

[3] Rothman LS, Gordon IE, Barbe A, Benner DC, Bernath PF, Birk M, et al. The HITRAN 2008 molecular spectroscopic database. J Quant Spectrosc Radiat Transfer 2009;110:533–72.

[4] Robert D, Bonamy J. Short range effects in semiclassical molecular line broadening calculations. J Phys (France) 1979;40:923–43.

[5] Lynch R, Gamache RR, Neshyba SP. Fully complex implementation of the Robert- Bonamy formalism: half widths and line shifts of H2O broadened by N2. J Chem Phys 1996;105:5711–21.

[6] Antony BK, Gamache PR, Szembek CD, D. Niles DL, Gamache RR. Modified complex Robert–Bonamy formalism calculations for strong to weak interacting systems. Mol Phys 2006;104:2791–9.

[7] Gamache RR, Laraia A. N2-, O2-, and air-broadened half-widths, their temperature dependence, and line shifts for the rotational band of $H2^{16}O$. J Mol Spec 2009;257:116–27.

[8] Jacquemart D, Gamache RR, Rothman LS. Semi-empirical calculation of air-broadened half-widths and air pressure-induced frequency shifts of water-vapor absorption lines. J Quant Spectrosc Radiat Transfer 2005;96:205–39.

[9] Ma Q, Tipping RH, Lavrentieva NN. Pair identity and smooth variation rules applicable for the spectroscopic parameters of H2O transitions involving high-*J* states. Mol Phys 2011;109:1925–41.

[10] Ma Q, Tipping RH, Lavrentieva NN. Theoretical studies of $N_2$-broadened half-widths of H2O lines involving high *J* states. Mol Phys 2012;110:307–31.

[11] Barber RJ, Tennyson J, Harris GH, Tolchenov RN. A high-accuracy computed water line list. Mon Not R Astron Soc 2006;368:1087–94.

[12] Tennyson J, Bernath PF, Brown LR, Campargue A, Császár AG, Daumont L, et al. IUPAC critical evaluation of the rotational-vibrational spectra of water vapor, Part III: Energy levels and transition wavenumbers for H216O. J Quant Spectrosc Radiat Transfer 2013;117:29–58.

[13] Tyuterev VG. The generating function approach to the formulation of the effective rotational Hamiltonian. J Mol Spectros 1992;151: 97–129.

[14] Mikhailenko SN, Tyuterev VG, Keppler KA, Winnewisser BP, Winnewisser M, Mellau G, et al. The $2\nu_2$ band of water: analysis of new FTS measurements and high-$K_a$ transitions and energy levels. J Mol Spectrosc 1997;184:330.

[15] Zou Q, Varanasi P. Laboratory measurement of the spectroscopic line parameters of water vapor in the 610–2100 and 3000–4050 $cm^{-1}$ regions at lower-tropospheric temperatures. J Quant Spectrosc Radiat Transfer 2003;82:45–98.

[16] Toth RA, Linelist of water vapor parameters from 500 to 8000 $cm^{-1}$, see ⟨http://mark4sun.jpl.nasa.gov/$H_2O$.html⟩.

[17] Jenouvier A, Daumont L, Régalia-Jarlot L, Tyuterev VG, Carleer M, A. Vandaele AC, et al. Fourier transform measurements of water vapor line parameters in the 4200–6600 $cm^{-1}$ region. J Quant Spectrosc Radiat Transfer 2007;105:326–55.