Recognition of They/Them as Singular Personal Pronouns in Coreference Resolution

Connor Baumler
University of Maryland, College Park
baumler@umd.edu

Rachel Rudinger
University of Maryland, College Park
rudinger@umd.edu

Abstract

As using they/them as personal pronouns becomes increasingly common in English, it is important that coreference resolution systems work as well for individuals who use personal "they" as they do for those who use gendered personal pronouns. We introduce a new benchmark for coreference resolution systems which evaluates singular personal "they" recognition. Using these *WinoNB schemas*, we evaluate a number of publicly available coreference resolution systems and confirm their bias toward resolving "they" pronouns as plural.

(1a) **The paramedic** performed CPR on the passenger even though *they* knew it was too late.
(1b) The paramedic performed CPR on **the passenger** even though *they* were already dead.
(2a) **The paramedics** tried to help Riley even though *they* knew it was too late.
(2b) The paramedics tried to help **Riley** even though *they* were already dead.

Figure 1: A (1) "Winogender" and corresponding (2) "WinoNB schema". The correct answers are bolded.

1 Introduction

While singular "they" has been widely used in the English-speaking world as a personal pronoun among nonbinary and other members of the LGBTQIA+ community for many years, it has taken time for this usage to be accepted by cisnormative society at large. In particular, professional institutions' acceptance of singular personal "they" has been much slower, but seen some uptake in recent years. The American Psychological Association began accepting singular "they" as a "self-identified pronoun" in 2019,[1] the same year Merriam-Webster added "nonbinary they" to their dictionary and made it their word of the year. [2]

As a result, English language models trained exclusively on sources such as old newspaper articles are likely not exposed to examples of singular "they" usage, in either the personal or generic setting. Even models trained on more recent data may miss this usage entirely if they rely on text that was subjected to conservative style guides. This means even state-of-the-art coreference resolution systems may not have recognition of singular "they."

In this work, we examine whether existing coreference resolutions systems are able to correctly resolve cases of singular personal "they". We find that when given a choice between resolving "they" correctly to a singular, named entity or to a group, current systems overwhelmingly choose to resolve "they" as plural or even choose not to resolve the pronoun at all. We also investigate these systems' recognition of singular generic "they" and find that this is a much more easily recognized use-case for some models. Overall, we find that models which have been trained on more contemporary or stylistically varied text that may contain examples of singular "they" to have the best performance in both the personal and generic case.

While failing to correctly resolve uses of singular personal "they" may feel like a trivial case to those who do not use it as their personal pronoun, this can cause a number of allocational and representational harms (Barocas et al., 2017; Blodgett et al., 2020; Cao and Daumé III, 2020; Dev et al., 2021). For example, a difference in coreference resolutions system performance between nonbinary people who use they/them pronouns and their peers who use binary personal pronouns would constitute a representational harm. This difference in performance can lead to allocational harms when coreference resolution is used in down-stream tasks

[1] Via APA Style
[2] Via Merriam-Webster

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3426 - 3432
July 10-15, 2022 ©2022 Association for Computational Linguistics

such as ranking authors based on citation counts in the bodies of texts (Dev et al., 2021).

2 Relevant Datasets

To evaluate coreference resolution systems' understanding of singular personal "they", we follow existing work (Rudinger et al., 2018; Zhao et al., 2018) that uses Winograd schemas to test for gender bias in such systems. The Winograd Schema Challenge (**WSC**) dataset consists of pairs of sentences in which there are two possible referents for a pronoun (Levesque et al., 2012). Based on a small edit, the pronoun in each sentence in a pair resolves to the opposite referent. These schemas' resolutions are designed to be obvious to a human reader but require deeper knowledge of the given situation for a model to understand them. Beyond the original WSC schemas, Rahman and Ng (2012) provide a set of definite pronoun resolution schemas (**DPR**) that focus on complex cases of definite pronouns that require world knowledge.

Multiple Winograd-style datasets exist to benchmark gender bias in coreference resolution systems.

Winogender schemas (Rudinger et al., 2018) use an occupation and a participant as their two possible referents. The correct resolution is clear from commonsense knowledge about what the person with the given occupation should be doing in the scenario. For example, in Figure 1, we know it doesn't make sense for the occupation referent "the paramedic" to be dead. We can see in the pair of paramedic examples how, by editing the circumstance (i.e., someone is already dead vs someone knows it is too late), we can change the correct resolution. These schemas are used to confirm that coreference resolution systems are more likely to choose interpretations that match with occupational gender stereotypes instead of the scenario.

WinoBias schemas (Zhao et al., 2018) are constructed similarly, though both possible referents are an occupation. These schemas are also split between cases in which the correct resolution can be found using purely syntactic information and cases that contain no syntactic clues, instead requiring deeper knowledge about the circumstance. Zhao et al. (2018) used these schemas to show that resolution systems will continue to make interpretations based on gender stereotypes, even when the correct answer can be chosen using only syntactic information.

Beyond the Winograd-style datasets we use in this work, there are a number of other coreference resolution datasets that focus on gender bias. GAP (Webster et al., 2018) consists of naturally-occurring ambiguous pronouns. It is balanced between male and female referents, but does not include instances of gender-neutral usage.

Cao and Daumé III (2020) investigate coreference resolution systems' ability to resolve gender neutral pronouns. Their work introduces two datasets: MAP and GICoref. MAP removes social gender cues such as gendered pronouns and semantically gendered nouns from the GAP dataset. They find that these changes dramatically decrease the accuracy of coreference systems. The GICoref dataset consists of naturally occurring text with examples of pronoun usage that are less common in prior datasets such as personal singular "they", neopronouns, and switching pronouns throughout a document. They find that coreference systems still have opportunity for improvement on this dataset, especially in the case of neopronouns. While their datasets contain ambiguous cases of singular personal "they", they do not explicitly test for coreferece resolution systems' understanding of the singular vs plural personal case. Our work is complementary to Cao and Daumé III (2020) as we focus on controlled experiments with constructed schemas (§3) rather than uncontrolled but naturally occurring text.

3 WinoNB Schemas

Winogender and Winobias schemas do not consider understanding of singular personal vs plural "they". While Winogender schemas do include sentences using singular "they", their two possible resolutions were both to individual referents. This means the tested systems had no choice to resolve "they" as a plural pronoun. Additionally, neither the occupation nor the participant referent in Winogender schemas is a specific, named person.

We would like to consider understanding of singular "they" when used as a personal pronoun. To do this, we modify appropriate schemas by hand from the Winogender, Winobias, WSC, and DPR datasets to create "WinoNB schemas".

To modify Winogender schemas into WinoNB schemas, we begin by changing the occupation referent to be a *group of people* and the participant to be a *named individual*. We will not use occupations to test for gender bias as Rudinger et al. (2018) did, but instead use them to make the scenario and the correct resolution more clear. In example 2a of Figure 1, it is sensible for the people with medical training to "know it was too late", but this could be more ambiguous if neither Riley nor the group explained to have medical training.

We further edit the schemas as necessary to ensure that the resulting sentences make sense. For instance, in Figure 1, we can see that it wouldn't have made sense for multiple paramedics to "perform CPR". Instead, we have the paramedics "try to help" the individual.

We also had many cases in which using "The [Occupation]s" was confusing as it sounded as though a group of people with the occupation were acting (or being acted upon) together. For instance, in "The physicians warned Riley that they needed to get more rest," it sounds as though a group of physicians are all giving Riley advice at once. In these cases, we change "The physicians" out for "Several physicians" so the sentence can be interpreted as Riley getting the advice from different physicians on multiple occasions.

We apply this same methodology to the WinoBias, WSC, and DPR datasets. We exclude examples with non-human referents (such as "The dog chased the cat, which ran up a tree. It waited at the {bottom/top}") and examples in which the scenario will not make sense with a singular and plural referent. This left 4077 templates that can be filled with individuals' names. The split between cases in which the pronoun should be interpreted as singular or plural was not perfectly even (with one more plural case than singular) as some datasets provided more than two possible predicates for a handful of scenarios. As our analysis will consider accuracy on the set of plural and singular cases individually, this slight imbalance will not affect the result.

The authors performed the edits to turn the existing Winogender, WinoBias, WSC, and DPR schemas into WinoNB schemas manually. A random subset of 100 of the templates (25 from each source) were verified by a fluent English speaker who resolved the pronouns with 96% accuracy.

3.1 Choice of Names

Since we are examining singular "they" as a personal pronoun, we will need to use the names of people in our examples. Due to biases in training data, pre-trained language models may not treat all names equally on downstream tasks (Shwartz et al., 2020).

To help account for this, we used 15 names to fill the individual's slot in each template. 10 of these were common AMAB (assigned male at birth) and AFAB (assigned female at birth) baby names, and 5 were baby names that were not strongly assigned to either (See Appendix A.1).

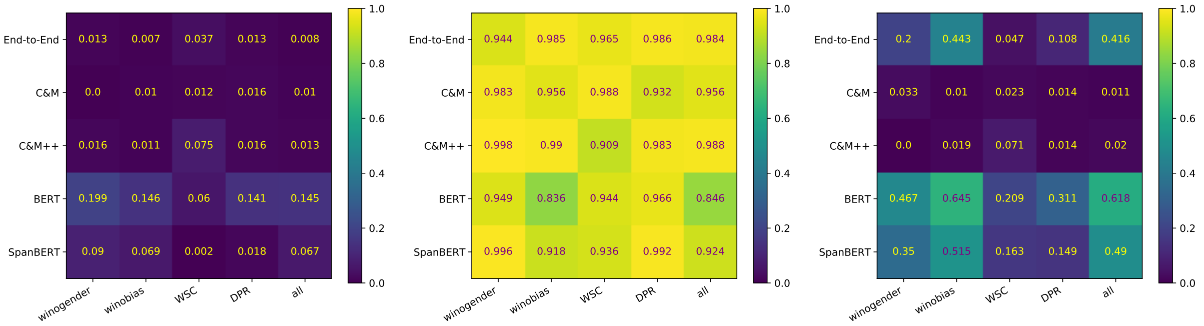
3.2 Using Singular Generic "They"

While some people find cases of singular personal "they" to be hard to resolve, singular generic "they" [is accepted by more people]. As Foertsch and Gernsbacher (1997) found, while singular "they" is less cognitively efficient than a binary pronoun when the gender of the referent is presumably known, generic singular "they" can be equally if not more cognitively efficient than a binary gendered pronoun.

To test if coreference resolution systems are more able to handle singular generic "they" than singular personal, we created an additional set of WinoNB schemas which use the generic "someone" instead of a named person for the singular referent. For instance, Example 2b in Figure 1 would be changed to "The paramedics tried to help **someone** even though *they* were already dead."

4 Results and Discussion

We evaluate a representative but not exhaustive set of five coreference resolution systems on our WinoNB schemas. First, we use Clark and Manning (2016)'s deep reinforcement learning system, which we will refer to as **C&M**. We call Lee et al. (2018)'s model with attention-based span representation refinement **End-to-End**. Both of these models were trained on the CoNLL-2012 dataset (Pradhan et al., 2012) which consists of coreference resolution problems in OntoNotes 5.0 (Weischedel et al., 2012). OntoNotes contains text from sources such as Newswire and magazines. We will also evaluate Hugging Face's model, which we call **C&M++**, which builds on C&M (Wolf, 2017). C&M++ was also trained on OntoNotes. Finally, we will evaluate **BERT** (Joshi et al., 2019) and **Span-BERT** (Joshi et al., 2020), whose original pre-training was done on the BooksCorpus (Zhu et al., 2015) and English Wikipedia with fine-tuning on OntoNotes.



(a) Singular Personal

(b) Plural

(c) Singular Generic

Figure 2: Model accuracies on WinoNB schemas containing different usages of “they”. These results are also shown as a table in Appendix A.2

4.1 Singular Personal “They” vs Plural “They”

First, we consider how well each of our models performs on the examples using singular personal vs plural “they”. As we can see in Figures 2a and 2b, the difference in performance is stark. On average, the models are 94.8% less likely to correctly resolve a WinoNB schema that uses singular personal “they” than one that uses plural “they”.

The different datasets did not exhibit a consistent ordering of difficulty across models. Using a Friedman Test (Friedman, 1937, 1940), we find that the ranking of the accuracies of each model over all datasets are not significantly different at the $p < .05$ level. However, using paired t-tests, we find that some models have significantly different performances ($p < .05$) on individual schema sources.

We find that BERT-base (which achieved about 3× the average accuracy on the singular personal case) has significantly better performance on three of the datasets, and its results are not significantly different from the other models on the dataset in which it underperforms. The three non-BERT-based models were all trained on OntoNotes 5.0, a dataset that is unlikely to contain instances of singular personal “they”. Sampling 100 sentences from OntoNotes 5.0 that contain a they/them pronoun, we found no cases of singular personal “they”.[3] While the BERT models were fine-tuned on OntoNotes, their pre-training on the BooksCorpus and Wikipedia may have contained singular personal “they” with Wikipedia officially allowing and encouraging the use of singular “they” since at least 2017.[4] From this, we may speculate that the BERT models’ relative success in handling singular personal “they” comes from exposure to the concept during pre-training.

While Span-BERT would have received similar exposure during pre-training, Span-BERT-base achieved a significantly lower accuracy than BERT-base on most WinoNB datasets. Since Joshi et al. (2020) reported that Span-BERT outperforms BERT on OntoNotes, we speculate that Span-BERT’s lower WinoNB accuracy comes from optimizing for OntoNotes, which may require a deprioritization of any knowledge of singular personal “they” gleaned from pre-training.

BERT’s increased accuracy may also have its drawbacks. While Figure 2a focuses solely on how many pronouns were correctly resolved as singular, many of the mistaken examples were either resolved to an unrelated entity or received no resolution at all. BERT-base failed to resolve 5.15% of singular personal “they” examples to the group or the individual. This rate is more than double the average rate of such failures over all models.

4.2 Singular Personal “They” vs Singular Generic “They”

Beyond cases of singular personal “they”, we also considered examples of singular generic “they” in which the individual referent is “someone”, not a named person. As we can see in Figure 2c, the

[3] We did find cases where of singular “they” referring to an instiution such as “Kraft Foods”. In many cases, the number of the pronoun was ambiguous given without larger context of the full document. As the OntoNotes entries from web data consist only of single sentences (Weischedel et al., 2012), some of these examples’ number could not be determined, but there was no evidence to suggest that any were intended to be singular, let alone singular personal.

[4] Via English Wikipedia’s gender-neutral language policy

models overall are much more able to handle singular generic "they", reaching an average 31.2% accuracy on these cases. The models were, on average, 6.4× more likely to correctly resolve a case of singular generic "they" versus singular personal.

4.3 Effects of Gendered Name Associations

By using a variety of names traditionally given to AFAB or AMAB babies or neither, we can investigate the effect of gender association of the individual's name on the models' willingness to choose the singular personal reading of "they". We do see some subtle differences in the results on these differently gendered names. For instance, BERT-base was about 7.5% more likely to incorrectly resolve Winogender-sourced schemas with traditionally AMAB names than AFAB or neutral names. These differences in performance were significant at the $p < .05$ level.

Overall, we find that in cases with significant difference in performance across differently gendered names, the models generally incorrectly resolve cases of singular personal "they" that ought to refer to masculine names. This could mean that these models more strongly associate common AMAB names with he/him pronouns than they do common AFAB names with she/her.

5 Conclusion

We have introduced "WinoNB schemas", a set of pronoun resolution pairs that test recognition of singular personal, singular generic, and plural "they" in English coreference resolution. These schemas are adapted from four existing sets of Winograd schemas. Testing on five publicly available off-the-shelf coreference models, we demonstrated that current models largely do not interpret "they" as a singular personal pronoun, though they are more likely to accept singular generic "they". We infer that this is due to popular training datasets largely containing text from times and settings in which singular personal "they" is unlikely to have been used. We find that BERT models, which may have seen cases of singular personal "they" during pretraining, are most able to solve WinoNB schemas.

WinoNB schemas can only demonstrate the existence of bias against nonbinary people, not its absence. Our methodology relies on "they" having both singular and plural usages, so it cannot be used to test for understanding of neopronouns such as xe/xem/xyr. A coreference resolution system can handle WinoNB schemas well and still perform poorly for members of the nonbinary community who use neopronouns or switch between multiple sets of pronouns. Still, these schemas can serve as a jumping-off point for creating and evaluating future models that better serve the nonbinary community.

References

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *SIGCIS Conference*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Julie Foertsch and Morton Ann Gernsbacher. 1997. In search of gender neutrality: Is singular they a cognitively efficient substitute for generic he? *Psychological science*, 8(2):106–111.

Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.

Milton Friedman. 1940. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2012. Ontonotes release 5.0.

Thomas Wolf. 2017. State-of-the-art neural coreference resolution for chatbots. Blog post.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Appendix

A.1 Names

We use name popularity data from 2013 that was collected from the Social Security Administration (SSA) and Five Thirty Eight, who also use this SSA data. Neutral names were chosen such that the gender ratio was not more than 7:3 in favor of either binary gender. Note that having a weak gender association as a baby name is neither a necessary nor sufficient condition for a name to be popular among nonbinary people. For our purposes, SSA data was the most readily available.

Gender Association	Name	% AFAB
fem	Sophia	99.88%
fem	Emma	99.91%
fem	Olivia	99.87%
fem	Isabella	99.90%
fem	Ava	99.89%
masc	Noah	0.44%
masc	Jacob	0.13%
masc	Liam	0.12%
masc	Mason	0.41%
masc	William	0.08%
neutral	Casey	41.57%
neutral	Riley	49.23%
neutral	Jessie	52.21%
neutral	Jackie	57.86%
neutral	Avery	66.47%

Table 1: Gendered names used to fill WinoNB templates.

		winogender	**winobias**	**WSC**	**DPR**	**all**
Singular Personal	End-to-End	0.013	0.007	0.037	0.013	0.008
	C&M	0.000	0.010	0.012	0.016	0.010
	C&M++	0.016	0.011	**0.075**	0.016	0.013
	BERT	**0.199**	**0.146**	0.060	**0.141**	**0.145**
	SpanBERT	0.090	0.069	0.002	0.018	0.067
Plural	End-to-End	0.944	0.985	0.965	0.986	0.984
	C&M	0.983	0.956	**0.988**	0.932	0.956
	C&M++	**0.998**	**0.990**	0.909	0.983	**0.988**
	BERT	0.949	0.836	0.944	0.966	0.846
	SpanBERT	0.996	0.918	0.936	**0.992**	0.924
Singular Generic	End-to-End	0.200	0.443	0.047	0.108	0.416
	C&M	0.033	0.010	0.023	0.014	0.011
	C&M++	0.000	0.019	0.071	0.014	0.020
	BERT	**0.467**	**0.645**	**0.209**	**0.311**	**0.618**
	SpanBERT	0.350	0.515	0.163	0.149	0.490

Table 2: Model accuracies on WinoNB schemas containing different usages of "they".

A.2 Reformatted Results

For readability, we include the same results from Figure 2 in Table 2.

A.3 Licence Information for Used Datasets and Models

Winogender (Rudinger et al., 2018), Winobias (Zhao et al., 2018), C&M (Clark and Manning, 2016), and C&M++ (Wolf, 2017) are released under MIT licenses. End-to-End (Lee et al., 2018), BERT (Joshi et al., 2019), and Span-BERT (Joshi et al., 2020) are released under Apache License 2.0. WSC (Levesque et al., 2012) are released under a Creative Commons Attribution 4.0 International License. We did not find license information for DPR (Rahman and Ng, 2012). Our use of these data and models does not conflict with the licenses' stated access conditions.