analytical chemistry

Article

pubs.acs.org/ac

Results of an Interlaboratory Comparison of Analytical Methods for Contaminants of Emerging Concern in Water

Brett J. Vanderford,*[,†] Jörg E. Drewes,[‡] Andrew Eaton,[§] Yingbo C. Guo,[∥] Ali Haghani,[§] Christiane Hoppe-Jones,[‡] Michael P. Schluesener,[⊗] Shane A. Snyder,[#] Thomas Ternes,[⊗] and Curtis J. Wood[$]

[†]Southern Nevada Water Authority, P.O. Box 99954, Las Vegas, Nevada 89193, United States

[‡]Colorado School of Mines, 1500 Illinois Street, Golden, Colorado 80401, United States

[§]Eurofins Eaton Analytical, Inc., 750 Royal Oaks Drive, Monrovia, California 91016, United States

[∥]Metropolitan Water District of Southern California, 700 Moreno Avenue, La Verne, California 91750, United States

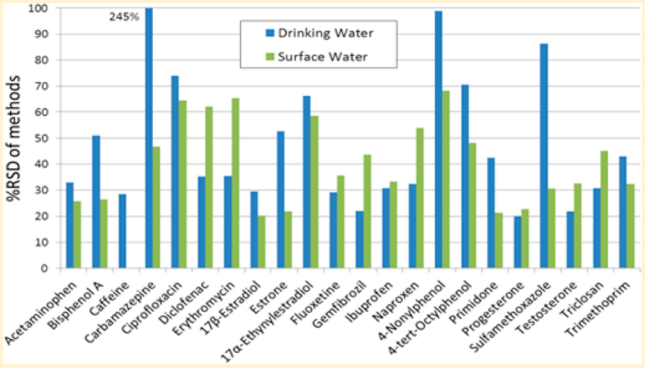
[⊗]German Federal Institute of Hydrology, Am Mainzer Tor 1, 56068 Koblenz, Germany

[#]University of Arizona, 1133 E. James E. Rogers Way, Tucson, Arizona 85721, United States

[$]Environmental Resource Associates, 6000 W. 54th Avenue, Arvada, Colorado 80002, United States

Supporting Information

ABSTRACT: An evaluation of existing analytical methods used to measure contaminants of emerging concern (CECs) was performed through an interlaboratory comparison involving 25 research and commercial laboratories. In total, 52 methods were used in the single-blind study to determine method accuracy and comparability for 22 target compounds, including pharmaceuticals, personal care products, and steroid hormones, all at ng/L levels in surface and drinking water. Method biases ranged from <10% to well over 100% in both matrixes, suggesting that while some methods are accurate, others can be considerably inaccurate. In addition, the number and degree of outliers identified suggest a high degree of variability may be present between methods currently in use. Three compounds, ciprofloxacin, 4-nonylphenol (NP), and 4-*tert*-octylphenol (OP), were especially difficult to measure accurately. While most compounds had overall false positive rates of ≤5%, bisphenol A, caffeine, NP, OP, and triclosan had false positive rates >15%. In addition, some methods reported false positives for 17β-estradiol and 17α-ethynylestradiol in unspiked drinking water and deionized water, respectively, at levels higher than published predicted no-effect concentrations for these compounds in the environment. False negative rates were also generally <5%; however, rates were higher for the steroid hormones and some of the more challenging compounds, such as ciprofloxacin. The elevated false positive/negative rates of some analytes emphasize the susceptibility of many current methods to blank contamination, misinterpretation of background interferences, and/or inappropriate setting of detection/quantification levels for analysis at low ng/L levels. The results of both comparisons were collectively assessed to identify parameters that resulted in the best overall method performance. Liquid chromatography−tandem mass spectrometry coupled with the calibration technique of isotope dilution were able to accurately quantify most compounds with an average bias of <10% for both matrixes. These findings suggest that this method of analysis is suitable at environmentally relevant levels for most of the compounds studied. This work underscores the need for robust, standardized analytical methods for CECs to improve data quality, increase comparability between studies, and help reduce false positive and false negative rates.



Contaminants of emerging concern (CECs), such as endocrine disrupting compounds (EDCs) and pharmaceuticals and personal care products (PPCPs), are organic contaminants that have been detected in wastewater (WW), surface water (SW), and drinking water (DW) throughout the world.[1−4] Their occurrence is most often the result of municipal wastewater discharge, as many of these compounds are not completely removed during treatment.[4] Because CECs are numerous and represent a broad spectrum of compounds, the development of techniques for their analysis is quite challenging. CECs vary widely in their physicochemical properties (e.g., polarity, molecular weight, pK_a, water

Received: October 10, 2013
Accepted: December 11, 2013
Published: December 11, 2013

   dx.doi.org/10.1021/ac403274a | *Anal. Chem.* 2014, 86, 774−782

solubility, etc.) and their concentrations in the environment can be quite low, typically subμg/L. As such, complex extraction and detection techniques are generally necessary. In addition, the widespread use of many target CECs make contamination of samples and equipment a common problem. Furthermore, matrix effects caused by nontarget, background interferences may result in improper data interpretation because the effects can vary substantially between matrixes and lead to the reporting of inaccurate concentrations.[5] At this time, no standardized, widely accepted methods exist for their analysis in water. These and other issues related to CEC analysis have led to the question of whether results generated by a given method accurately depict the true concentration of each contaminant in water and whether the results from various methods are comparable.

The objectives of this study were to investigate the ability of a large number of methods being used in ongoing research and monitoring efforts to accurately quantify a target group of PPCPs using interlaboratory comparisons and assess the comparability of the reported data on a compound-by-compound basis. Several reports of interlaboratory comparisons involving some of the target compounds have been published in the literature. The majority of the comparisons focused on a limited number of compounds with varying numbers of laboratories. Ternes et al.[6] carried out a study on carbamazepine and diclofenac in groundwater (GW) and SW with three laboratories but only to confirm the quality of the methods. Sengl and Krezmer[7] focused mainly on analgesics and β-blockers with between 11 and 20 laboratories while Farre et al.[8] conducted their comparison on nonsteroidal anti-inflammatory drugs and 11−14 laboratories. Another paper investigated three steroids in GW, SW, and WW with between 8 and 12 laboratories.[9] Finally, Drewes et al.[10] studied a large group of compounds with three laboratories; however, no study has examined a large number of diverse compounds with a similarly large number of participating laboratories and their methods.

This paper presents the results of two single-blind interlaboratory comparisons performed in DW and SW. In total, 25 research and commercial laboratories participated in the study using a total of 52 methods to analyze 22 CECs, including a number of PPCPs and suspected EDCs. The accuracy and comparability of the methods, the rates of false positives and false negatives, and the factors resulting in the best overall method performance are discussed.

EXPERIMENTAL SECTION

Target Compound Selection. Because of the large number of CECs that have been detected in water, the number of potential compounds had to be narrowed considerably. While the target compound list had to be limited for obvious practical reasons, this meant that the selection criteria needed to produce a list that was both widely acceptable and not overwhelming in terms of logistical demands. Therefore the target compound list was carefully developed using a number of different factors, including the potential compounds' physicochemical properties (e.g., polarity, functional groups, molecular weights, etc.), their degree of use/sales, their published occurrence (frequently detected was preferable) and fate (recalcitrant or easily degraded) both in the environment and during DW treatment, their occurrence on lists published by various research groups and internationally recognized research organizations such as the Global Water Research Coalition,[11] and their status as an unregulated, emerging contaminant for which there were no standardized or rigorously validated methods. These factors led to the target compound list shown in Table 1.

Table 1. Target Compounds

PPCPs		potential EDCs	
compound	CAS no.	compound	CAS no.
acetaminophen	103-90-2	bisphenol A	80-05-7
caffeine	58-08-02	17β-estradiol	50-28-2
carbamazepine	298-46-4	estrone	53-16-7
ciprofloxacin	86393-32-0	17α-ethynylestradiol	57-63-6
diclofenac	15307-79-6	4-nonylphenol	84852-15-3
erythromycin	114-07-8	4-*tert*-octylphenol	140-66-9
fluoxetine	56296-78-7	progesterone	57-83-0
gemfibrozil	25812-30-0	testosterone	58-22-0
ibuprofen	15687-27-1		
naproxen	22204-53-1		
primidone	125-33-7		
sulfamethoxazole	723-46-6		
triclosan	3380-34-5		
trimethoprim	738-70-5		

Participating Laboratories/Methods. In total, 25 laboratories participated in the two interlaboratory comparisons using a total of 52 methods. The laboratories were located in the United States, Canada, Europe, and Australia and routinely employed the methods for testing purposes at the time of the comparisons. The laboratories were allowed to use their own method(s) of choice and were not asked to alter their method(s) for the study. A summary of the methods is given in Table 2, and the details of each method are presented in the Supporting Information (Tables S1 and S2).

Sample Preparation and Scheme. The first interlaboratory comparison used finished DW from the River Mountains Water Treatment Facility (300 mgd) in Henderson, Nevada, which treats water from Lake Mead, NV. The second used

Table 2. Summary of Methods

category	methods	category	methods
Laboratory/Analyst Experience		Extraction	
<12 months	16	SPE	47
>12 months	36	liquid/liquid extraction	3
		online SPE	2
Instrument		Calibration	
GC/MS	8	external	2
LC−MS	1	internal	7
LC−MS/MS	40	isotope dilution	25
online LC−MS/MS	2	combinations	18
LC−TOF-MS	1		
Method Basis			
in-house	14		
EPA 1694[a]	4		
EPA 539	1		
published literature[a]	29		
other	4		

[a]With and without modifications.

dx.doi.org/10.1021/ac403274a | *Anal. Chem.* 2014, 86, 774−782

Table 3. Spike Levels (Assigned Values) For Interlaboratory Comparisons (ng/L)

	both comparisons		DW comparison			SW comparison	
target compounds	DI water and unspiked DW/SW	low DI spike	low DW spike	high DI spike	high DW spike	low DI and SW spikes	high SW spikes
acetaminophen	0	12.2	13.5	68.8	62.5	10.0	65.0
bisphenol A	0	13.0	14.5	98.9	110	23.0	59.9
caffeine	0	19.3	17.5	68.8	62.5	N/A[a]	N/A[a]
carbamazepine	0	8.77	7.97	87.7	79.7	16.2	54.8
ciprofloxacin	0	31.9	29.0	60.5	55.0	45.0	275
diclofenac	0	16.7	18.5	56.3	62.5	14.0	43.0
erythromycin	0	13.5	15.0	93.5	85.0	14.0	50.0
17β-estradiol	0	4.50	5.00	18.0	20.0	2.88	12.5
estrone	0	4.95	5.50	19.3	17.5	2.75	12.5
17α-ethynylestradiol	0	5.50	5.00	26.4	24.0	2.75	12.5
fluoxetine	0	8.41	9.35	73.4	66.8	5.12	34.7
gemfibrozil	0	7.65	8.50	60.5	55.0	5.75	31.0
ibuprofen	0	9.90	11.0	67.5	75.0	18.0	55.0
naproxen	0	13.8	12.5	74.3	67.5	11.0	65.0
4-nonylphenol	0	105	95.0	605	550	105	550
4-*tert*-octylphenol	0	24.7	27.5	659	599	55.0	270
primidone	0	8.10	9.00	49.5	55.0	22.5	47.0
progesterone	0	4.95	5.50	20.4	18.5	2.50	13.0
sulfamethoxazole	0	7.15	6.50	99.0	90.0	29.0	75.0
testosterone	0	4.50	5.00	20.3	22.5	3.13	12.5
triclosan	0	9.45	10.5	76.5	85.0	18.0	50.0
trimethoprim	0	8.10	9.00	68.8	62.5	12.0	44.0

[a]N/A: Not applicable.

WW-influenced SW (~4% WW) from the Las Vegas Bay of Lake Mead, Nevada. Water quality data from the two sites can be found in the Supporting Information (Tables S3 and S4). Water was collected from these locations in 100 L drums, quenched with ascorbic acid (DW only) and preserved with sodium azide (1 g/L), as it has been shown that these agents have very little impact on the target compounds.[12] Controls were prepared in deionized (DI) water. The water was then shipped to a commercial proficiency testing provider for homogenization and preparation of the test samples.

Stock solutions of the target analytes were prepared by Restek Corporation (State College, PA). These standards were used to prepare the test samples and also were sent to the participating laboratories for use as calibration standards. This was done to remove one source of variability among the laboratories and allowed the methods to be more effectively evaluated without the confounding factors of standard stability or different sources of standards. Appropriate amounts of intermediate solutions prepared from the stock solutions were spiked into the samples to yield the targeted spike concentrations, as discussed below.

The DW comparison consisted of the following eight samples: one unspiked DI water blank, one unspiked DW sample, one "low level" DI water spike (4.50−105 ng/L), two duplicate "low level" DW spikes (5.00−95.0 ng/L), one "high level" DI water spike (18.0−659 ng/L), and two duplicate "high level" DW spikes (17.5−599 ng/L). The SW comparison consisted of one unspiked DI water blank, two duplicate unspiked SW samples, one "low level" DI water spike (2.5 − 105 ng/L), two duplicate "low level" SW spikes (2.5−105 ng/L), and two duplicate "high level" SW spikes (12.5−550 ng/L).

In both comparisons, the target compounds were randomly placed into three separate groups (Supporting Information Tables S5 and S6) and spiked according to the matrix in Matrixes S1 and S2 in the Supporting Information. In this manner, there were no samples of exclusively one type, but every analyte had a representative of each type of sample in the comparison. Spike values for each comparison are shown in Table 3. The spike values were set low enough to challenge the methods but not so low that less than seven methods would be able to detect a given analyte (based on the method reporting limits (MRLs) reported by the participants).

Calculations and Analysis. Percent bias was calculated according to the following equation:

$$\text{percent bias} = \left(\frac{C_r - C_u}{C_s} - 1\right) \times 100 \quad (1)$$

where C_r = concentration of the target analyte reported by the method, C_u = concentration in the unspiked sample, and C_s = assigned concentration for the target analyte in the sample.

For this study, C_s was considered to be the spiked value and, for most analytes in most samples, $C_u = 0$. As such, percent bias for a given sample/analyte was calculated by simply dividing the reported concentration (C_r) by C_s, subtracting 1 and multiplying by 100. However, in the SW comparison, ambient levels of some of the target compounds were observed in the unspiked SW samples. Thus, to calculate percent bias, these levels needed to be taken into account. For methods that reported a detectable concentration in the unspiked SW samples that was not accompanied by significant hits in the deionized water blanks, C_u was defined as the average of the concentrations in the unspiked samples (nondetects were not included in this calculation). If there were also a significant number of detections in the accompanying DI water blanks, the detections in the unspiked SW were deemed unreliable and were not used to define C_u. In addition, consensus concentrations were calculated by determining the median of the reported concentrations of all the laboratories for each of the unspiked samples. If the MRL of a method was greater than

dx.doi.org/10.1021/ac403274a | Anal. Chem. 2014, 86, 774−782

Table 4. Bias by Compound in DW Interlaboratory Comparison (%)

target compound	no. of methods	bias range	overall median bias	DI spike median bias	low DW spike median bias	high DW spike median bias	overall DW spike median bias
acetaminophen	21	2.5−123	16	15	14	22	17
bisphenol A	18	3.3−206	35	33	47	23	37
caffeine	25	4.2−138	19	22	32	16	21
carbamazepine	28	1.5−2720	9.1	7.9	9.6	12	9.5
ciprofloxacin	14	19−488	41	37	40	58	49
diclofenac	19	5.1−97	17	20	19	18	18
erythromycin	16	4.8−79	21	18	20	24	26
17β-estradiol	19	1.7−65	18	16	18	7.7	17
estrone	18	2.5−241	14	15	15	9.1	16
17α-ethynylestradiol	21	3.8−374	17	13	15	16	18
fluoxetine	19	2.2−66	18	24	19	15	17
gemfibrozil	23	3.7−81	11	9.4	11	14	12
ibuprofen	27	1.7−101	16	11	21	15	18
naproxen	24	2.7−120	17	16	14	13	14
4-nonylphenol	11	15−739	51	88	54	42	43
4-*tert*-octylphenol	13	6.0−249	41	41	41	21	43
primidone	15	2.2−185	16	9.1	25	13	19
progesterone	14	3.7−36	15	13	14	14	14
sulfamethoxazole	25	1.8−506	13	12	11	13	13
testosterone	15	2.4−49	13	18	13	9.8	12
triclosan	23	2.2−59	19	20	20	13	21
trimethoprim	23	2.7−126	18	16	17	13	16

Table 5. Bias by Compound in SW Interlaboratory Comparison (%)

target compound	no. of methods	bias range	overall median bias	DI spike median bias	low SW spike median bias	high SW spike median bias	overall SW spike median bias
acetaminophen	19	4.2−90	18	12	14	10	16
bisphenol A	17	5.0−100	23	28	27	16	22
carbamazepine	26	2.6−186	13	10	12	13	14
ciprofloxacin	12	13−1001	59	67	52	39	49
diclofenac	20	3.8−375	24	16	31	16	28
erythromycin	15	4.6−261	24	23	23	21	22
17β-estradiol	19	6.7−35	18	14	17	19	18
estrone	19	4.1−38	20	9.1	24	18	18
17α-ethynylestradiol	21	9.5−264	19	26	31	18	19
fluoxetine	18	4.5−149	26	33	34	16	34
gemfibrozil	21	2.9−225	11	6.4	9.3	8.1	11
ibuprofen	24	2.0−152	11	9.4	12	9.7	11
naproxen	22	1.6−253	14	11	14	10	11
4-nonylphenol	10	12−415	38	74	45	22	30
4-*tert*-octylphenol	14	7.4−415	31	26	30	35	31
primidone	15	3.7−85	6.4	4.9	8.1	6.5	13
progesterone	14	6.2−41	15	7.4	12	15	15
sulfamethoxazole	23	2.0−192	10	6.9	13	8.7	12
testosterone	16	4.0−71	11	13	10	8.0	12
triclosan	21	4.4−340	17	19	18	14	18
trimethoprim	21	4.2−74	14	8.3	17	15	14

the consensus concentration for a given compound (i.e., a method reported a nondetect for the unspiked SW samples), C_u was defined as the average of the consensus concentrations of the two unspiked SW samples. If a method reported nondetects for the unspiked SW samples and the MRL was less than the consensus concentration (i.e., a false negative), C_u was set to zero.

The absolute value of the percent bias on each sample was used to calculate an average absolute bias (henceforth referred to as "bias", for clarity) on each compound for each method. The absolute value was used to prevent positive and negative bias from canceling each other out, thereby masking the degree of bias present. After careful consideration, it was decided that the complexities of the involvement of false negative nondetects in the bias calculations far outweighed their impact on the conclusions of the study and were therefore not considered in the calculation of absolute bias. The median bias of all the

dx.doi.org/10.1021/ac403274a | *Anal. Chem.* 2014, 86, 774−782

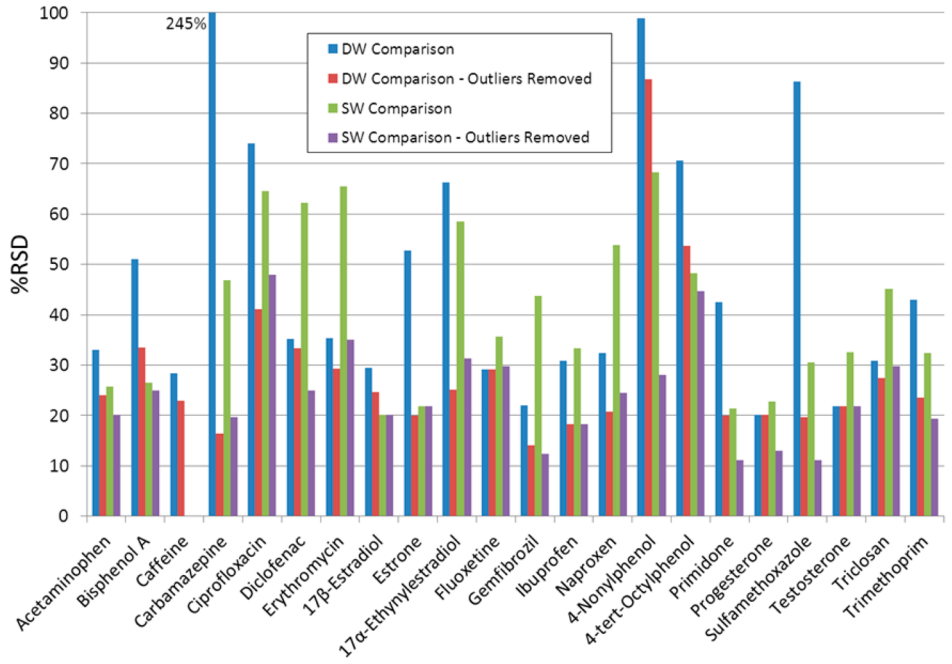


Figure 1. % RSDs of each compound for DW and SW comparisons. *DW comparison % RSD for carbamazepine is 245%.

methods used for each compound could then be used to compare two different compounds in terms of relative bias.

To calculate the overall percent relative standard deviation (% RSD) for each compound, the % RSD of each spiked sample was calculated and then the average of the % RSDs from the six spiked samples was determined for each comparison. In addition, possible outliers were identified for each target compound using the Grubb's Test at the 95% confidence level. Outliers were identified on a sample by sample basis. A second % RSD and average % RSD were calculated after the stepwise removal of any outliers identified using this statistical test. Nondetects (including false negatives) were not considered in this analysis. % RSDs were also calculated for the unspiked SW samples that had ambient concentrations of the target compounds that were detectable by the majority of the methods and determined by comparison with accompanying DI water blanks to not be the result of blank contamination.

All statistical calculations were performed using Statgraphics Centurion XV (Statpoint, Inc., Warrenton, VA), version 15.2.12.

Sample Logistics. Samples were shipped to the participating laboratories in coolers at 4 °C using overnight delivery. Instructions were included with the shipment stating that the samples were ready for extraction/analysis as received. In addition, directions were provided for reporting the results. Laboratories were sent enough of each sample to cover the extraction volume reported in their method(s). A deadline was given for reporting data, and only data received prior to the deadline were considered in the study.

RESULTS AND DISCUSSION

Two interlaboratory comparisons, one in DW and one in SW, were performed in this study. To assess method performance on a compound-by-compound basis, the average absolute bias for each method was calculated. The median absolute bias of the methods, for each compound, was found to be a helpful tool and was used in the data analysis, as discussed below, because of the ease with which it allowed comparison on a compound-by-compound basis. It also significantly reduced the effects of outliers. Results of the comparisons are summarized in Tables 4 and 5.

Performance by Compound in DW Comparison. One compound, carbamazepine, had an overall median bias of <10% and the majority of compounds (12 of 22) had median biases between 15 and 20%, indicating the results for the majority of methods for most compounds were within 20% of the assigned value. However, four compounds, bisphenol A (BPA), ciprofloxacin, 4-nonylphenol (NP), and 4-*tert*-octylphenol (OP), had median biases that were much greater than the others at 35%, 41%, 51%, and 41%, respectively, indicating these compounds were more challenging, in general, for the methods to analyze.

Bias ranges indicated that all compounds had at least one method with a large degree of bias. Only one compound had a maximum bias of <40% (progesterone), and many compounds had methods with biases of at least 100%. Bias ranges also indicated that most compounds had at least one method with a small degree of bias. All but two compounds had at least one method with a bias <10% demonstrating that accurate methods are available for most compounds. Ciprofloxacin and NP were the only particularly poor exceptions with all methods having median biases of >19% and >15%, respectively.

When broken down into DI versus DW median bias, many of the compounds were similar (<15% difference). This may indicate that the DW matrix did not play a drastic role in the performance of the methods. A total of 15 of the 22 compounds had higher median biases in DW than in DI water, but of those 15, only 5 (ciprofloxacin, 17α-ethynylestradiol (EE2), erythromycin, ibuprofen, and primidone) increased by more than 30%, suggesting that these compounds may be particularly sensitive to matrix effects. The median biases of the remaining seven compounds either were the same or were reduced in DW. Of those seven, two (NP and testosterone) were reduced by more than 30%.

The difference between median biases for the low versus the high DW spikes is shown in Table 4. The median biases of 14

dx.doi.org/10.1021/ac403274a | *Anal. Chem.* 2014, 86, 774−782

of 22 compounds were less in the high DW spikes. In addition, 9 of the 22 compounds decreased more than 30% while only two increased by more than 30%, showing that it was more difficult for the methods to quantify the target compounds at the lower spike levels.

Performance by Compound in SW Comparison. Only one compound, primidone, had an overall median bias of <10% and the majority of compounds (13 of 21) had median biases between 10 and 20%. One compound, ciprofloxacin, had a median bias >50%, indicating that half of the methods had an average bias of over 50% from the assigned value. NP and OP had similarly high median biases of 31% and 38%, respectively, showing large discrepancies between the assigned and reported values for these compounds. All compounds had at least one method with a bias of <10% with the exceptions of ciprofloxacin and NP. On the other hand, many compounds (14) had at least one method report >100% bias. This indicates that although there were many methods that had a high degree of bias, there also were methods that were able to accurately quantify most of the target compounds in SW.

When split into DI versus SW median bias, more than half of the compounds showed higher median bias in the SW than in the DI water with 9 compounds >30% higher. Seven compounds had a higher median bias in the DI matrix than in the SW matrix but of those, only four (BPA, ciprofloxacin, EE2 and NP) were >10% higher. The difference between DI and SW median bias was particularly pronounced for diclofenac, estrone (E1), gemfibrozil, primidone, progesterone, and trimethoprim, suggesting that the methods for these compounds were especially sensitive to SW matrix effects.

The vast majority (17 of 21) of median biases were lower in the high SW spikes than in the low SW spikes. Of those, 14 showed a decrease of >15%. As the second comparison was designed to challenge the methods with a higher degree of matrix interferences and lower spike amounts, this was not unexpected and agrees with the literature.[7−9] Nevertheless, this demonstrates the increased inaccuracy of methods for these compounds at low ng/L levels.

Method Comparability. To quantify the degree to which the methods agreed on the concentrations of the target compounds in the spiked samples, the average % RSD was calculated for each target compound from the results of the DW and SW comparisons, as discussed in the Experimental Section. Because of the significant impact that a small number of values had on the % RSD for a number of the compounds, outliers were identified and removed and % RSDs were recalculated. Figure 1 shows the % RSDs with and without outliers for each compound in both matrixes.

Similar to previous work,[8] many of the target compounds quantified by various methods were determined to have outliers in both matrixes. Only fluoxetine, progesterone, and testosterone in the DW comparison and 17β-estradiol (E2) and E1 in the SW comparison did not have outliers for at least one of the samples. For example, in the DW comparison, the average % RSD of carbamazepine decreased from 245% to 16% due to the removal of a few outliers from methods that had high degrees of bias (Performance Factor Results Tables in the Supporting Information). Sulfamethoxazole also had a large decrease in % RSD from 86% to 20%, largely due to one method that had all of its data points determined as outliers. For the SW comparison, several compounds had much lower average % RSDs when the outliers were removed. Erythromycin decreased from 65% to 35% after all of one method's results were removed and naproxen dropped from 54% to 24% after all of the SW results from two methods and all of the low spike SW results from another method were removed. NP dropped significantly from 68% to 28% after a number of low spike outliers were removed and the % RSDs for diclofenac and gemfibrozil were also lowered from 62% to 25% and 44% to 12%, respectively, after outliers were removed.

Overall, the majority of analytes had % RSDs between 20% and 30% for both comparisons, after outliers were removed, with the median % RSD for DW and SW being 24% and 22%, respectively. Six compounds (carbamazepine, gemfibrozil, ibuprofen, progesterone, primidone, and sulfamethoxazole) had % RSDs <20% in both matrixes, indicating the methods for these compounds had the most agreement. In contrast, ciprofloxacin had % RSDs in DW and SW of 41% and 48%, respectively. With outliers included, these values ballooned to 74% and 64%. In addition, the two surfactant-related compounds NP and OP had % RSDs >50% in at least one of the comparisons and >70% with outliers included. The number and degree of outliers identified in this study suggest that large discrepancies may exist between values given by various methods in the literature due to a lack of standardization. In particular, data comparability for ciprofloxacin, NP, and OP may be especially difficult.

Ambient concentrations in the unspiked SW also provided an opportunity to assess method comparability and were determined for eight compounds: carbamazepine, gemfibrozil, NP, OP, primidone, sulfamethoxazole, triclosan, and trimethoprim. As with the spiked samples, extreme outliers were present for all but one of the compounds (Supporting Information Table S7). While median values between the two unspiked samples were remarkably close, % RSDs were inconsistent and very high because of extreme outliers. % RSDs ranged from 38% to 148% with some methods reporting values that were often 3−10+ times greater than the median value for that sample. After removing the outliers, the % RSDs were much more consistent and ranged from 11 to 34% for the eight compounds, with the exception of OP with a % RSD of 72% for one of the SW samples.

Among results reported by individual methods, most methods had % RSD values of <5%, regardless of the matrix (data not shown). Overall, the median % RSD of the methods for each compound were <10% (Supporting Information Table S8) showing that individual methods were precise, which agrees with another study.[8] Exceptions to this were low spikes for BPA (DW), E2 (DW), NP (DW/SW), OP (DW/SW), and triclosan (DW) and both SW spikes for ciprofloxacin. In addition, low spikes in both matrixes generally led to higher % RSDs than high spikes, indicating that within-method variability increased as spike concentrations decreased.

False Positive and False Negative Analysis. False positives (FPs) and false negatives (FNs) were examined to determine how frequently values were reported for unspiked samples and how many values were reported as nondetect when, in fact, the samples had been spiked. FPs may be related to blank contamination and FNs can be due to degradation of the target analyte prior to analysis. Both FPs and FNs may also be due to the setting of detection/quantification limits that overestimate the ability of the method to accurately quantify a target compound at that level. For the DW comparison, the blank samples consisted of one DI water blank and one unspiked DW sample. For the SW comparison, there was one DI water blank and two unspiked SW samples. As some of the

dx.doi.org/10.1021/ac403274a | Anal. Chem. 2014, 86, 774−782

Table 6. Summary of False Positives and False Negatives from Both Interlaboratory Comparisons

compound	no. false positives	no. of samples	rate (%)	range (ng/L)	no. of false negatives	no. of samples	rate (%)
acetaminophen	1	99	1	1.3	9	221	4
bisphenol A	21	87	24	1.2−46	2	193	1
caffeine	8	50	16	2.93−29.2	N/A[a]	N/A[a]	N/A[a]
carbamazepine	4	82	5	2.01−24.4	5	298	2
ciprofloxacin	9	64	14	12−112	17	144	12
diclofenac	3	98	3	1.4−4.99	0	214	0
erythromycin	1	47	2	26.4	4	171	2
17β-estradiol	2	95	2	2.27−5	22	209	11
estrone	3	93	3	1.55−2.62	15	203	7
17α-ethynylestradiol	5	105	5	2.6−13.8	23	231	10
fluoxetine	1	92	1	1.5	15	204	7
gemfibrozil	0	67	0	N/A[a]	5	243	2
ibuprofen	14	126	11	1−33	3	282	1
naproxen	6	114	5	3.2−18.9	5	254	2
4-nonylphenol	22	32	69	12.8−1280	11	116	9
4-*tert*-octylphenol	9	40	23	1.6−130	3	148	2
primidone	3	45	7	1.43−22.6	1	165	1
progesterone	2	70	3	1.54−1.66	7	154	5
sulfamethoxazole	4	73	5	2.88−5.17	0	265	0
testosterone	0	78	0	N/A[a]	15	170	9
triclosan	13	67	19	1.28−350	4	243	2
trimethoprim	0	67	0	N/A[a]	4	243	2

[a]N/A: Not applicable.

compounds had ambient concentrations in the SW, these detections were taken into account when the methods were assessed for FPs/FNs. A summary of the FPs and FNs for the two comparisons is presented in Table 6, and results for the individual comparisons can be found in the Supporting Information Tables S9 and S10.

Overall, most compounds had FP rates of ≤5%. Ibuprofen, a frequently used anti-inflammatory agent, and caffeine and triclosan, two commonly used personal care products, had overall FP rates of 11%, 16%, and 19%, respectively, most likely due to blank contamination. Ciprofloxacin had a FP rate of 14% while BPA had a rate of 24%. NP had the highest rate of FPs at 69%, including 64% in the DW comparison and 80% in the SW comparison. Although the rate of FPs was low (<5%), the concentrations of E2 and EE2 reported in the DW and SW blanks ranged from 2.3 to 5.0 ng/L and 2.6 to 14 ng/L, respectively. These values are higher than the predicted no-effect concentrations of 2.0 ng/L for E2 and 0.1 ng/L for EE2, as reported in the literature for ecological health,[13] and reflect a similar instance of false positives reported by the United States Geological Survey in conjunction with their study published in 2001.[14]

In general, FN rates were lower than FP rates. Only three compounds had FN rates ≥10%. Again, ciprofloxacin proved challenging with a FN rate of 12%. Overall, two of the steroids (E2 and EE2) had FN rates >10% with the remaining three (E1, progesterone and testosterone) having rates between 5−9%. All steroid FN rates increased in the SW comparison, most likely because the spike levels were lowered for this round and the matrix was more complex. These results agree with previous work that suggests steroidal analysis may be difficult at concentrations ∼5 ng/L,[9] depending on whether single- or triple-quadrupole instruments were used. The elevated false positive and false negative rates of some analytes underscore the susceptibility of samples to blank contamination and/or improper determination of detection/quantification levels for analysis at low ng/L levels.

These results demonstrate the urgent need for standardized analytical methods for CECs that are capable of accurate measurements at typical ambient levels and also show the value of including blind spikes in existing monitoring programs. While most CEC methods used in this study appear to be relatively precise, they are not necessarily accurate and the level of bias can be extreme.

Determination of Performance Factors. The results of both comparisons were collectively assessed to identify parameters that resulted in the best overall method performance (i.e., <10% overall bias). Only methods that reported results for both comparisons were used in the assessment. In addition, methods that did not report values for more than half of the samples in either round were not included. This situation typically arose when the MRL of a method was not sensitive enough to detect the low spike samples. The biases of each method for both comparisons were averaged to obtain an overall bias. The methods were then ranked from least to greatest bias on a compound by compound basis and compared against each method's reported parameters to identify trends. Assessed parameters included laboratory experience (<12 months vs >12 months), analyst experience (<12 months vs >12 months), extraction technique, instrument, calibration technique, and ionization mode. All data are shown in the Performance Factor Results Tables section of the Supporting Information.

In contrast with previous studies,[7,10] no trends or correlations were found between overall bias and experience or extraction technique; however, methods employing LC−MS/MS were able to achieve overall biases of <10% for most compounds (Table 7), while GC/MS methods were consistently in the lower 50th percentile. These conclusions agree with a previous interlaboratory comparison study[9] but disagree with two others,[7,8] although one of these studies found that

dx.doi.org/10.1021/ac403274a | Anal. Chem. 2014, 86, 774−782

Table 7. Recommended Analysis Techniques, By Compound, Based on Frequency in Methods with <10% Overall Bias

compound	ionization mode	analysis	calibration
acetaminophen	ESI+	LC−MS/MS	isotope dilution
bisphenol A	no conclusion	no conclusion	no conclusion
caffeine	ESI+	LC−MS/MS	isotope dilution
carbamazepine	ESI+	LC−MS/MS	isotope dilution
ciprofloxacin	no conclusion	no conclusion	no conclusion
diclofenac	ESI−	LC−MS/MS	isotope dilution
erythromycin	ESI+	LC−MS/MS	isotope dilution
17β-estradiol	APCI+	LC−MS/MS	isotope dilution
estrone	ESI− or APCI+	LC−MS/MS	isotope dilution or internal calibration (pre-SPE)
17α-ethynylestradiol	no conclusion	no conclusion	no conclusion
fluoxetine	ESI+	LC−MS/MS	isotope dilution
gemfibrozil	ESI−	LC−MS/MS	isotope dilution
ibuprofen	ESI−	LC−MS/MS	isotope dilution
naproxen	ESI−	LC−MS/MS	isotope dilution or internal calibration
4-nonylphenol	no conclusion	no conclusion	no conclusion
4-*tert*-octylphenol	no conclusion	no conclusion	no conclusion
primidone	ESI+	LC−MS/MS	isotope dilution
progesterone	ESI+ or APCI+	LC−MS/MS	isotope dilution
sulfamethoxazole	ESI+	LC−MS/MS	isotope dilution
testosterone	ESI+ or APCI+	LC−MS/MS	isotope dilution
triclosan	no conclusion	no conclusion	no conclusion
trimethoprim	ESI+	LC−MS/MS	isotope dilution

GC/MS led to the presence of more outliers.[8] In most cases, a single ionization technique was shared by all methods within a particular instrumental class (e.g., all EI for GC/MS vs all ESI− for LC−MS/MS), with the notable exception being the steroids. The ionization techniques most frequently used in the methods with the least overall biases are given in Table 7.

Isotope dilution as a calibration technique also consistently led to low degrees of bias for many of the compounds. Seven compounds (caffeine, carbamazepine, erythromycin, primidone, progesterone, sulfamethoxazole, and trimethoprim) had biases of <10% only with methods that used isotope dilution (with or without extraction of the calibration curve). In addition, isotope dilution was used in the method with the least overall bias and the majority of methods with <10% biases for diclofenac, gemfibrozil, ibuprofen, and testosterone. Two other compounds, E1 and naproxen, had more than one calibration technique able to consistently provide <10% overall bias. Three compounds (BPA, ciprofloxacin, and OP) could not be accurately analyzed on a consistent basis by any of the methods in this study and three other compounds (EE2, OP, triclosan) could only be analyzed by one method with <10% overall bias. In these cases, no technique was recommended.

CONCLUSIONS

Because of the high degree of variability between methods and the number and degree of outliers observed, results from this study clearly demonstrate that robust, standardized methods are necessary for many prominent CECs currently being discussed in the literature and under consideration for regulatory action. Hundreds of papers have been published on the analysis of CECs, yielding a wealth of information on such topics as occurrence and fate in natural and engineered environments and the potential for adverse impacts to environmental and human health. However, these results suggest that data comparability in published literature regarding CECs may be difficult, especially among water matrixes (wastewater, surface water, drinking water, etc.) and even between waters of the same matrix.

To improve low level CEC determination, it is recommended that particular attention be paid to blank contamination and reporting limit determination to reduce the rates of false negatives and false positives. Furthermore, it is suggested that analysis via LC−MS/MS coupled with the calibration technique of isotope dilution strongly be considered during the development of any standardized method for these compounds.

Perhaps most importantly, results from this work likely suggest that some studies in the literature have very high degrees of analytical bias and/or large numbers of false positives/negatives. Further, the use of occurrence data from unsuitable analytical procedures may have resulted in inappropriate risk assessments and prioritization for regulation. Thus, it is important that the consequences these data potentially have had on past decisions is recognized and critical that analytical quality and reliability be considered in future assessments.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: brett.vanderford@snwa.com. Phone: (702) 856-3659. Fax: (702) 856-3647.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was funded in part by the Water Research Foundation (Project No. 4167), a joint owner of certain technical information upon which this publication was based. The comments and views detailed herein may not necessarily reflect the views of the Water Research Foundation, its officers, directors, affiliates, or agents. The authors thank the Analytical Research and Development group at the Southern Nevada Water Authority for their valuable efforts.

REFERENCES

(1) Kolpin, D. W.; Furlong, E. T.; Meyer, M. T.; Thurman, E. M.; Zaugg, S. D.; Barber, L. B.; Buxton, H. T. *Environ. Sci. Technol.* **2002**, *36* (6), 1202−1211.

(2) Boyd, G. R.; Reemtsma, H.; Grimm, D. A.; Mitra, S. *Sci. Total Environ.* **2003**, *311*, 135−149.

(3) Benotti, M. J.; Trenholm, R. A.; Vanderford, B. J.; Holady, J. C.; Stanford, B. J.; Snyder, S. A. *Environ. Sci. Technol.* **2009**, *43* (3), 597−603.

(4) Ternes, T. A.; Stumpf, M.; Mueller, J.; Haberer, K.; Wilken, R.-D.; Servos, M. *Sci. Total Environ.* **1999**, *225*, 81−90.

(5) Vanderford, B. J.; Snyder, S. A. *Environ. Sci. Technol.* **2006**, *40*, 7312−7320.

(6) Ternes, T. A.; Meisenheimer, M.; McDowell, D.; Sacher, F.; Brauch, H.-J.; Haist-Gulde, B.; Preuss, G.; Wilme, U.; Zulei-Seibert, N. *Environ. Sci. Technol.* **2002**, *36*, 3855−3863.

(7) Sengl, M.; Krezmer, S. *Accred. Qual. Assur.* **2003**, *8* (11), 523−529.
(8) Farre, M.; Petrovic, M.; Gros, M.; Kosjek, T.; Martinez, E.; Heath, E.; Osvald, P.; Loos, R.; Le Menach, K.; Budzinski, H.; De Alencastro, F.; Mueller, J.; Knepper, T.; Fink, G.; Ternes, T. A.; Zuccato, E.; Kormali, P.; Gans, O.; Rodil, R.; Quintana, J. B.; Pastori, F.; Gentili, A.; Barceló, D. *Talanta* **2008**, *76*, 580−590.
(9) Esperanza, M.; Herry, G.; Manciot, F.; Laine, J. M. *Water Practice Technol.* **2006**, *1* (2), 1−8.
(10) Drewes, J.; Sedlak, D.; Snyder, S.; Dickenson, E. *Development of Indicators and Surrogates for Chemical Contaminant Removal during Wastewater Treatment and Reclamation*; WateReuse Foundation: Alexandria, VA, 2008.
(11) de Voogt, P.; Janex-Habibi, M.-L.; Sacher, F.; Puijker, L.; Mons, M. *Water Sci. Technol.* **2009**, *59* (1), 39−46.
(12) Vanderford, B. J.; Mawhinney, D. B.; Trenholm, R. A.; Zeigler-Holady, J. C.; Snyder, S. A. *Anal. Bioanal. Chem.* **2011**, *399*, 2227−2234.
(13) Caldwell, D. J.; Mastrocco, F.; Anderson, P. D.; Lange, R.; Sumpter, J. P. *Environ. Toxicol. Chem.* **2012**, *31* (6), 1396−1406.
(14) USGS Toxic Substances Hydrology Program. ERRATA for "Pharmaceuticals, Hormones, and Other Organic Wastewater Contaminants in U.S. Streams, 1999−2000: A National Reconnaissance"; Environmental Science & Technology, v. 36, no. 6, pp 1202−1211. http://toxics.usgs.gov/regional/est_errata.html (accessed December 5, 2013).