A&A 640, A67 (2020)
https://doi.org/10.1051/0004-6361/202037450

Astronomy & Astrophysics

Simulating JWST deep extragalactic imaging surveys and physical parameter recovery

O. B. Kauffmann[1], O. Le Fèvre[1], O. Ilbert[1], J. Chevallard[2], C. C. Williams[3], E. Curtis-Lake[4,5], L. Colina[6,7], P. G. Pérez-González[6], J. P. Pye[8], and K. I. Caputi[9]

[1] Aix Marseille Univ., CNRS, LAM, Laboratoire d'Astrophysique de Marseille, Marseille, France
e-mail: olivier.kauffmann@lam.fr
[2] Sorbonne Universités, UPMC-CNRS, UMR7095, Institut d'Astrophysique de Paris, 75014 Paris, France
[3] Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721, USA
[4] Cavendish Astrophysics, University of Cambridge, Cambridge CB3 0HE, UK
[5] Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
[6] Centro de Astrobiología, Departamento de Astrofísica, CSIC-INTA, Cra. de Ajalvir km.4, 28850 Torrejón de Ardoz, Madrid, Spain
[7] International Associate, Cosmic Dawn Center (DAWN) at the Niels Bohr Institute, University of Copenhagen and DTU-Space, Technical University of Denmark, Copenhagen, Denmark
[8] University of Leicester, School of Physics & Astronomy, Leicester LE1 7RH, UK
[9] Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands

Received 6 January 2020 / Accepted 2 June 2020

ABSTRACT

We present a new prospective analysis of deep multi-band imaging with the *James Webb* Space Telescope (JWST). In this work, we investigate the recovery of high-redshift $5 < z < 12$ galaxies through extensive image simulations of accepted JWST programs, including the Early Release Science in the EGS field and the Guaranteed Time Observations in the HUDF. We introduced complete samples of ~300 000 galaxies with stellar masses of $\log(M_*/M_\odot) > 6$ and redshifts of $0 < z < 15$, as well as galactic stars, into realistic mock NIRCam, MIRI, and HST images to properly describe the impact of source blending. We extracted the photometry of the detected sources, as in real images, and estimated the physical properties of galaxies through spectral energy distribution fitting. We find that the photometric redshifts are primarily limited by the availability of blue-band and near-infrared medium-band imaging. The stellar masses and star formation rates are recovered within 0.25 and 0.3 dex, respectively, for galaxies with accurate photometric redshifts. Brown dwarfs contaminating the $z > 5$ galaxy samples can be reduced to <0.01 arcmin^{-2} with a limited impact on galaxy completeness. We investigate multiple high-redshift galaxy selection techniques and find that the best compromise between completeness and purity at $5 < z < 10$ using the full redshift posterior probability distributions. In the EGS field, the galaxy completeness remains higher than 50% at magnitudes $m_{\rm UV} < 27.5$ and at all redshifts, and the purity is maintained above 80 and 60% at $z \leq 7$ and 10, respectively. The faint-end slope of the galaxy UV luminosity function is recovered with a precision of 0.1–0.25, and the cosmic star formation rate density within 0.1 dex. We argue in favor of additional observing programs covering larger areas to better constrain the bright end.

Key words. galaxies: high-redshift – galaxies: photometry – galaxies: distances and redshifts – galaxies: fundamental parameters – galaxies: evolution

1. Introduction

The detection of distant sources has mainly been driven by multi-wavelength photometry, through deep imaging over selected areas of the sky. The *Hubble* Space Telescope (HST), with its Advanced Camera for Surveys (ACS) and Wide-Field Camera 3 (WFC3), enabled the discovery of many high-redshift galaxies with its deep optical and near-infrared (IR) imaging (e.g., Scoville et al. 2007; Koekemoer et al. 2007; Wilkins et al. 2011; Schenker et al. 2013; McLure et al. 2013; Bouwens et al. 2015), effectively covering the rest-frame ultraviolet (UV) region of these sources. Near-infrared observations are necessary to detect high-redshift galaxies because of the strong attenuation blueward of the Lyman limit by the intergalactic medium (IGM), and as the Universe becomes more neutral, the flux that is blueward of Lyman alpha also gets attenuated. The mid-infrared observations with the *Spitzer* Space Telescope have improved the characterization of galaxy physical properties that are required to constrain galaxy evolution from the epoch of reionization to the present day (e.g., Sanders et al. 2007; Caputi et al. 2015). In particular, *Spitzer* has provided most of the constraints on the rest-frame optical at high redshift (Oesch et al. 2014, 2018). The census of high-redshift sources is particularly important to estimate which sources contributed most of the ionizing photons necessary to support neutral hydrogen reionization. The latest accounts point to a high number of faint sources producing enough ionising photons (Bouwens et al. 2015; Robertson et al. 2015), which reconcile a late reionization supported by the latest cosmic microwave background (CMB) constraints on the Thomson scattering optical depth (Planck Collaboration VI 2020) and UV photons from galaxy counts (Madau & Dickinson 2014). Establishing a complete and unbiased census of galaxies and associated ionizing photons remains a priority in order to understand this important transition phase in the Universe, which is directly linked to the formation of the first galaxies.

Identifying high-redshift galaxies within, and following, the epoch of reionization ($5 < z < 12$) is challenging because of their low number density which decreases with redshift. The methods to select high-redshift candidates mostly rely on the identification of the dropout in continuum emission blueward of Lyman alpha (Steidel et al. 1996). Lyman break galaxies (LBG) can be identified through color-color selections, mainly using photometry in the rest-frame UV. Alternatively, photometric redshifts obtained from spectral energy distribution (SED) fitting make use of all the photometric information (e.g., McLure et al. 2013; Finkelstein et al. 2015), spanning the optical to near-infrared, but they do introduce model dependencies. With the large number of low-redshift sources, which are several orders of magnitude more numerous, the high-redshift galaxy samples are subject to contamination because of the similar colors of these sources in the observed frame. The main contaminants are low-redshift ($z \sim 1-2$) dust-obscured galaxies with very faint continuum in the visible bands (Tilvi et al. 2013). Brown dwarfs are other potential contaminants of the $z > 5$ galaxy samples because of their similar near-infrared colors. The number of detected sources increases with telescope sensitivity, which naturally leads to an increasing probability of finding multiple objects along the line-of-sight. Therefore, the impact of source blending becomes more important (Dawson et al. 2016). In the case of source confusion, the background estimation becomes more challenging and individual sources are harder to isolate. The background level by itself also affects source separation, so that extended sources with internal structures may be mistaken for multiple nearby objects. In addition, the galaxy morphology is more complex at high redshifts (Ribeiro et al. 2016), therefore requiring adapted source detection techniques.

The *James Webb* Space Telescope (JWST[1], Gardner et al. 2006), which is to be launched in 2021, will revolutionize near- and mid-infrared astronomy. It will provide the first subarcsecond high-sensitivity space imaging ever at wavelengths above 3 microns and up to 25 microns, overcoming the current limitations of ground-based and space-based observatories. The onboard instruments include two imaging cameras, the Near-Infrared Camera (NIRCam[2], Rieke et al. 2005), and the Mid-Infrared Instrument (MIRI[3], Rieke et al. 2015; Wright et al. 2015), which together cover the wavelength range from 0.6 to 28 microns. These capabilities are perfectly suited for the discovery and the study of high-redshift galaxies during the epoch of reionization at $z > 6$, in combination with the deep optical imaging from HST and other ancillary data.

Predictions are required for preparation of the deep JWST imaging programs. The observed number counts per field of view and their redshift distribution need to be quantified, as well as the source detectability and the completeness and purity of the selected samples, depending on the detection method. The most direct number count predictions require the integral of the luminosity function multiplied by the differential comoving volume over a given area and redshift interval. High-redshift luminosity functions may be estimated by either extrapolating some lower-redshift measurements or using semi-analytic modeling (Mason et al. 2015a; Furlanetto et al. 2017; Cowley et al. 2018; Williams et al. 2018; Yung et al. 2019).

These methods quantify the expected number of detectable sources in a given field, not the number of sources which may be extracted and correctly characterized. Alternatively, the recovery of the galaxy physical parameters may be simulated with mock galaxy photometry and SED-fitting procedures. Bisigello et al. (2016) tested the derivation of galaxy photometric redshifts with JWST broad-band imaging, considering multiple combinations of NIRCam, MIRI and ancillary optical bands. The galaxy physical parameter recovery was investigated using the same methodology (Bisigello et al. 2017, 2019). Analogously, Kemp et al. (2019) analyzed of the posterior constraints on the physical properties from SED-fitting with JWST and HST imaging.

The aim of this paper is to investigate how to best identify high-redshift galaxies in the redshift range $5 < z < 12$ from JWST deep-field imaging, to estimate their number counts, with associated completeness and purity, and how their physical parameters can be recovered, focusing on stellar mass (M_*) and star formation rate (SFR). We concentrate on the identification and characterization of high-redshift sources from photometry, which will be required to identify sources for spectroscopic follow-up with JWST (NIRSpec, Birkmann et al. 2016). The simulation of deep fields necessitates the construction of realistic mock samples of sources, including all galaxies at all redshifts, as well as stars from the Galaxy. Any contamination estimate relies on the ability to produce simulations with sources which have realistic distributions of physical properties as a function of redshift, including fluxes and shapes projected on the image plane, as currently documented. In determining magnitudes, we need to include emission lines with strength corresponding to what is actually observed. In this way the contamination of high-redshift galaxy samples by low-redshift interlopers and Galactic stars can be estimated. We neglect quasars and transient objects. Existing observations are not deep enough to use as a basis for predictions for JWST and therefore some extrapolations are needed. To take geometrical effects into account, we generate mock images from the current knowledge of the instruments, then extract and identify sources. This allows us to more realistically characterize the statistical properties of the galaxy population, and especially source blending, thanks to the complete source sample. Figure 1 summarizes our methodology to make our forecasts.

This paper is organized as follows. In Sect. 2 we present the mock source samples, including galaxies and stellar objects. Section 3 describes our methodology to simulate images, extract sources and measure photometry and physical parameters. The results of the physical parameter recovery are detailed in Sect. 4. Section 5 describes our source selection investigations, including the rejection of the stellar contaminants, high-redshift galaxy selection and luminosity function computation. We summarize and conclude in Sect. 6. Magnitudes are given the AB system (Oke 1974), and we adopt the standard ΛCDM cosmology with $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$ and $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$.

2. Mock source samples

2.1. Mock galaxy sample

We build our galaxy sample from the JADES extraGalactic Ultradeep Artificial Realizations v1.2 (JAGUAR[4], Williams et al. 2018) developed for the JWST Advanced Deep Extragalactic Survey (JADES). This phenomenological model of galaxy evolution generates mock galaxy catalogs with physical and morphological parameters, reproducing observed statistical functions. Publicly available realizations consist of complete samples of star-forming and quiescent galaxies with stellar mass

[1] http://www.stsci.edu/jwst
[2] https://jwst-docs.stsci.edu/near-infrared-camera
[3] https://jwst-docs.stsci.edu/mid-infrared-instrument
[4] https://neogal.iap.fr/JAGUAR_mock_catalogue/

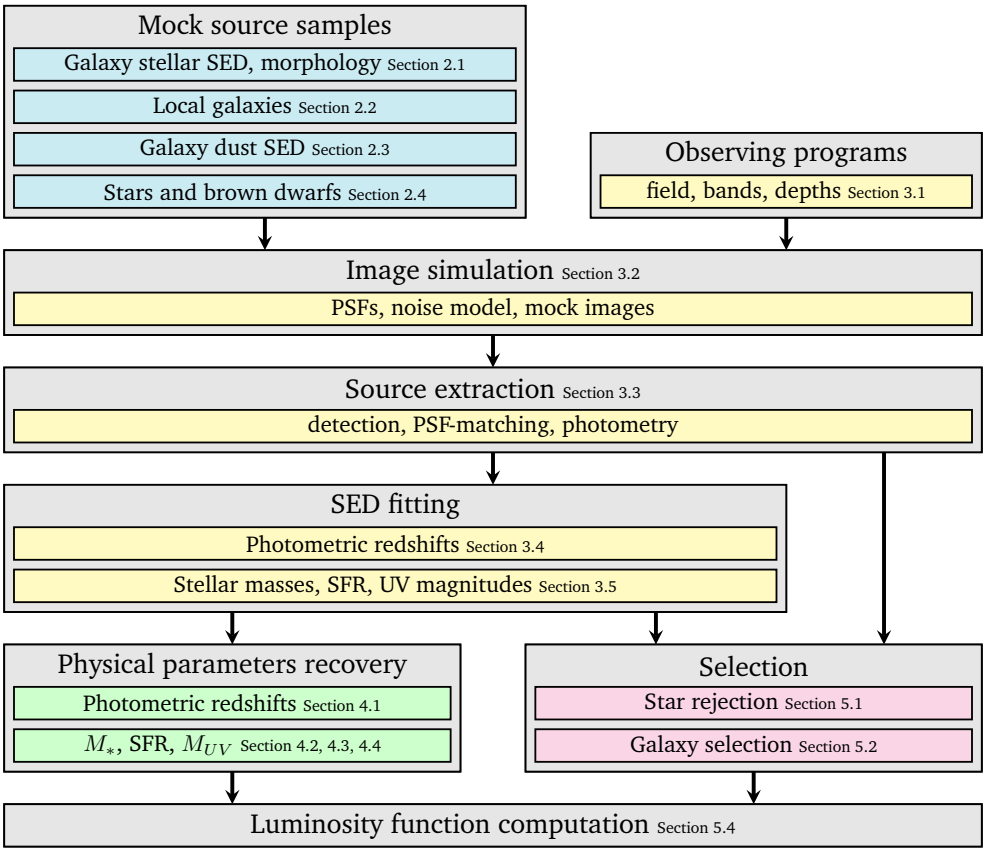

Fig. 1. Diagram summarizing the procedures to make our predictions. The gray boxes indicate the essential steps, and colored boxes show the detail of the subsections. The colors code for the main sections.

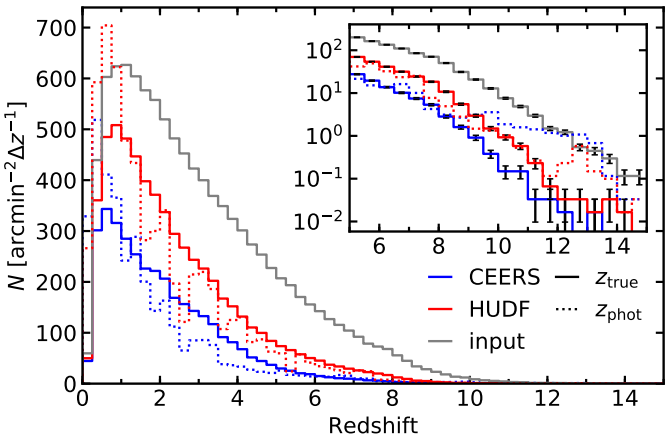

Fig. 2. Galaxy surface number density versus redshift. The gray line includes all the mock galaxies with stellar masses $\log(M_*/M_\odot) > 6$. The colored lines illustrate the redshift distribution of the detected sources in our CEERS and HUDF simulations. The solid and dashed lines represent input and photometric redshift distributions, respectively. The inset provides a zoom-in at high redshift, with Poisson error bars.

$6 < \log(M_*/M_\odot) < 11.5$ and redshift $0.2 < z < 15$ on areas of 11×11 arcmin^2, each containing $\sim 3 \times 10^5$ sources.

Stellar masses and redshifts are sampled from a continuous stellar mass function (SMF) model, constructed from the empirical SMF constraints of Tomczak et al. (2014) at $z < 4$ and the luminosity functions (LF) of Bouwens et al. (2015) and Oesch et al. (2018) at $z < 10$. We note that these observations support a rapid evolution of the UVLF at $z > 8$ inducing a strong decrease in galaxy number counts, which is still debated in the literature (e.g., McLeod et al. 2015, 2016). The SMF model separately describes star-forming and quiescent galaxies and is extrapolated beyond $z = 10$. Physical parameters (e.g., UV magnitude $M_{\rm UV}$, UV spectral slope β) are sampled from observed relationships and their scatter between $M_{\rm UV}$ and M_*, and $M_{\rm UV}$ and β. A spectral energy distribution (SED) is assigned to each set of physical parameters using BEAGLE (Chevallard & Charlot 2016). Williams et al. (2018) describe a galaxy star formation history (SFH) with a delayed exponential function and model stellar emission with the latest version of the Bruzual & Charlot (2003, hereafter BC03) population synthesis code. They consider the (line and continuum) emission from gas photoionized by young, massive stars using the models of Gutkin et al. (2016). Dust attenuation is described by the two-component model of Charlot & Fall (2000) and parametrized in terms of the $V-$band attenuation optical depth $\hat{\tau}_V$ and the fraction of attenuation arising from the diffuse ISM (set to $\mu = 0.4$), while IGM absorption follows the prescriptions of Inoue et al. (2014). No galaxies composed of metal-free population III stars are considered because of the lack of knowledge about these objects. No active galactic nuclei (AGN) models are considered either. Figure 2 shows the redshift distribution of the mock galaxy catalog.

Galaxy morphology is parametrized by one Sérsic function (Sérsic 1963) assumed to be wavelength-independent. Effective radii are sampled from a continuously evolving model with stellar mass and redshift, based on the observed size-mass relations in CANDELS and the 3D-HST survey by van der Wel et al. (2014). This model separately treats star-forming and quiescent galaxies, and extrapolates the observed trends down to $\log(M_*/M_\odot) = 6$. Axis ratio and Sérsic indices are sampled from the redshift-dependent distributions of van der Wel et al. (2012). The description of galaxy surface brightness profiles

relies on observed UV magnitudes and apparent shape measurements, so that only the strong lensing shape distortions are neglected here. Magnification is naturally included, although we do not expect magnification bias to be important in JWST pencil-beam surveys containing few $M_{\rm UV} < -22$ galaxies (Mason et al. 2015b). The underlying assumptions on the morphology of galaxies, especially at $z > 2$, may have important consequences on whether a source can be recovered, with two main limitations. Size measurements which assume symmetry find that sizes decrease with redshift (Shibuya et al. 2015, although cf. Curtis-Lake et al. 2016), whereas one finds that sizes remain large and constant with redshift when more adapted isophotal limits are used (Ribeiro et al. 2016). This arises from galaxies becoming more complex, multi-component as redshift increases, therefore spreading the total flux over a large area with surface brightness becoming lower. In addition, a clumpy galaxy may be resolved at certain wavelengths but appear mono-component in others because of the change of intrinsic structure and/or the varying angular resolution. This could be an increasing problem with source identification and multi-wavelength photometry. While this work is based on symmetric profiles, we will investigate how multi-component galaxies can be detected in a future paper.

Galaxy coordinates are sampled from a uniform distribution over the surveyed area. We therefore neglect galaxy alignment from clustering or lensing. The position of galaxy pairs with large line-of-sight separations are independent, meaning that the blending of high-redshift galaxies with low-redshift sources should remain unchanged with and without clustering, at the first order. Multiple over-dense regions may happen to be on the same line of sight, however we neglect these cases. Moreover, we only simulate nonconfused, high-resolution imaging, so that we expect our predictions to be negligibly impacted by clustering.

Because of the rarity of the high-redshift sources, large areas of the sky need to be simulated to make predictions with sufficient statistical significance. We replicate three times the initial galaxy catalog covering $11 \times 11\,\mathrm{arcmin}^2$ to increase the simulated area and sample size. For each replication, stellar masses are sampled from a centered log-normal distribution with $\sigma = 0.1$ dex. This ensures that the difference in the resulting SMF remains below ∼10% at $M_* < 10^{11}\,M_\odot$. Fluxes and SFRs are modified accordingly. Coordinates and galaxy position angles are randomized, and the other parameters kept unchanged.

2.2. Local galaxies

Low-redshift galaxies are one of the main sources of contamination for the high-redshift galaxy samples, notably because of the degeneracy between the Lyman and the Balmer breaks (e.g., Le Fèvre et al. 2015). Pencil-beam surveys contain very few local galaxies, however the apparent size of these objects are the largest on the sky, so that it is important to take them into account when simulating realistic blending.

The JAGUAR galaxy catalog does not include $0 < z < 0.2$ galaxies, because of the lack of low-redshift volume considered in building the stellar mass function in Tomczak et al. (2014). We sample redshifts and stellar masses at $M_* > 10^6\,M_\odot$ from the SMF continuous model of Wright et al. (2018) to fill this redshift interval. In that paper, the authors made use of the GAMA ($60\,\mathrm{deg}^2$), G10-COSMOS ($1\,\mathrm{deg}^2$) and 3D-HST ($0.274\,\mathrm{deg}^2$) data set gathered by Driver et al. (2018) to efficiently constrain both the bright and the faint ends of the SMF. For comparison, the area of the data used in Tomczak et al. (2014) is $316\,\mathrm{arcmin}^2$. We sample about 460 (50) galaxies with $\log(M_*/M_\odot) > 6$ (8) over $11 \times 11\,\mathrm{arcmin}^2$. We assign the spectrum from the JAGUAR $0.2 < z < 0.4$ galaxy with the closest stellar mass to each sampled parameter set. The maximum stellar age in these galaxies is therefore underestimated. By construction, about half of these galaxies are quiescent. The morphological parameters are sampled from the same distributions as in JAGUAR.

2.3. Galaxy infrared spectra

Dust emission can make a significant contribution to the near- and mid-infrared galaxy spectrum. In addition, mid-infrared photometry may considerably help to identify low-redshift contaminants to high-redshift samples using photometric redshift estimation (e.g., Ilbert et al. 2009). Because the galaxy spectra in JAGUAR include stellar and nebular emission, we include the additional dust emission for a more accurate modeling of the galaxy mid-infrared spectra. We neglect dust emission for low-mass[5] quiescent galaxies because Williams et al. (2018) neglected dust attenuation for these objects.

We take the library of dust spectral energy distributions of Schreiber et al. (2018) constructed from the dust models of Galliano et al. (2011). These templates separately describe the dust grain continuum emission and the polycyclic aromatic hydrocarbon (PAH) emission. The contribution of an AGN torus to the dust emission is neglected. The dust temperature ($T_{\rm dust}$) determines the shape of both components, the mid-to-total infrared color (IR8 = $L_{\rm IR}/L_{8\mu\rm m}$) sets their relative contributions and the infrared luminosity ($L_{\rm IR}$) scales the sum.

We attribute $T_{\rm dust}$ and IR8 to all the mock galaxies following the empirical laws evolving with redshift from Schreiber et al. (2018), including the intrinsic scatter. These relations were calibrated from the stacked *Spitzer* and *Herschel* photometry (Schreiber et al. 2015). We estimate the infrared luminosities from the $V-$band attenuation optical depth $\hat{\tau}_V$, assuming that the absorbed flux is entirely re-emitted by the dust (energy balance). We neglect the birth clouds component of the Charlot & Fall (2000) attenuation curve, since the JAGUAR catalog only provides the summed emission from young and old stars. This may lead to underestimated dust emission, as well as the limitation to the diffuse ISM. Figure 3 indicates a better agreement between simulated and empirical counts in the MIRI/$F770W$ filter.

2.4. Mock star sample

In this section, we present our formalism to generate mock stars from the Milky Way in the field of view. The strategy to create the mock star catalog is the following: (1) estimate the number density per spectral type, (2) sample heliocentric distances and physical properties, then (3) assign the spectrum with the closest properties.

We make use of the Besançon Model of the Galaxy[6] (Robin et al. 2003, 2012, 2014) to generate mock stars of spectral type FGKM. This model of stellar population synthesis provides star samples with intrinsic parameters (mass, age, metallicity, effective temperature $T_{\rm eff}$, surface gravity $\log g$). OBA stars are not sampled because of their rarity in pencil-beam surveys. We follow the galaxy model of Caballero et al. (2008) to determine the mean number of LTY stars per unit area. The galaxy density profile is modeled by an exponential thin-disk with the parameters from Chen et al. (2001), reliable

[5] With $\log(M_*/M_\odot) < 8.7 + 0.4z$.
[6] http://model2016.obs-besancon.fr/

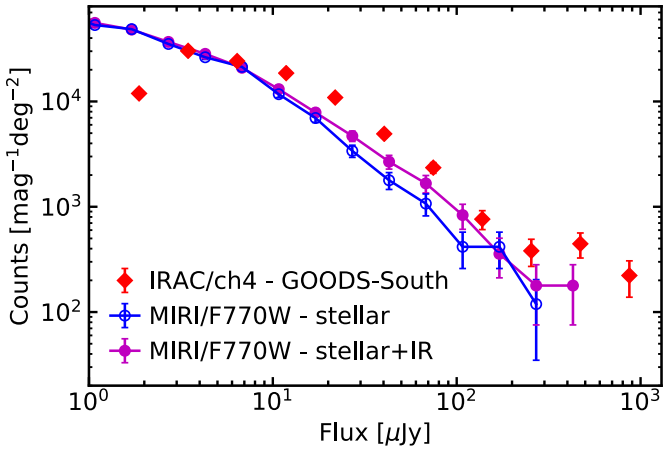

Fig. 3. Differential galaxy number counts in the MIRI/*F770W* filter with and without the dust emission, compared to the *Spitzer* IRAC/8 μm number counts measured in the GOODS-South field (Schreiber et al. 2017).

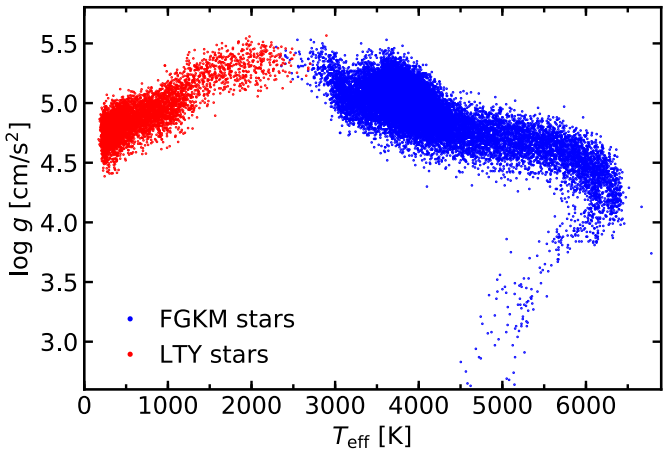

Fig. 4. Distribution of effective temperature and surface gravity for the mock FGKM stars from the Besançon model in blue, and our mock LTY stars in red.

at high galactic latitudes b. The surface density of objects at the central galactic coordinates (l, b) results from the integration of the density profile over heliocentric distance, scaled to the local number density. We take the predicted local number densities of Burgasser (2007) for L0 to T8 stars (see Caballero et al. 2008). Because of the small number of Y star observations, their number density is poorly constrained so we linearly extrapolate the local number densities of hotter stars to the cooler subtypes T9 and Y0-Y2. Star coordinates are sampled from a uniform distribution.

An effective temperature $T_{\rm eff}$ is assigned to each stellar subtype following the brown dwarf compilation[7] from Pecaut & Mamajek (2013). Surface gravity $\log g$ is computed from the stellar masses and radii, the latter being taken from the same compilation but imposing the lower bound of 0.1 $R_\odot$ (1 Jupiter radius). Stellar masses come from a linear model with 0.1 $M_\odot$ for type L0 and 0.02 $M_\odot$ for Y2. We include a bivariate Gaussian scatter to the duplet ($T_{\rm eff}$, $\log g$) with 10% relative dispersion. Figure 4 shows the sampled parameters for LTY stars. We sample the sedimentation efficiencies $f_{\rm sed}$ from Gaussian random distributions with mean $\mu = 2$ and scatter $\sigma = 1$ for L types, and

[7] http://www.pas.rochester.edu/~emamajek/

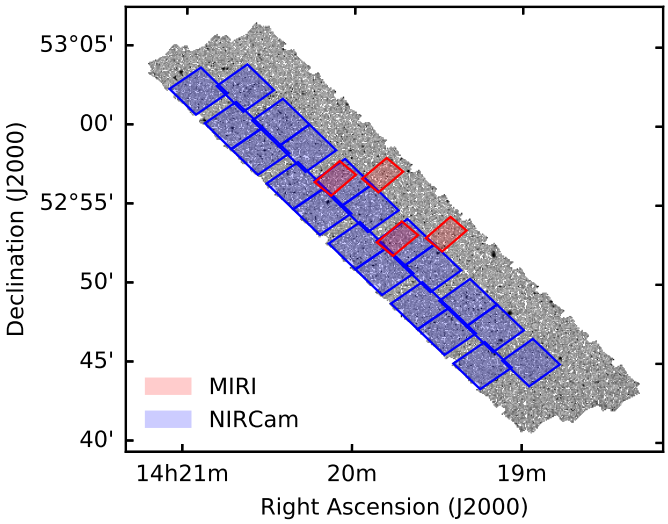

Fig. 5. CEERS layout in the EGS field. The 10 NIRcam imaging pointings are shown in blue and the 4 MIRI parallels in red. The ancillary HST/WFC3 H_{160}-band coverage is in gray. The pointings are all approximate until the final schedule. The parallel NIRSpec observations are not represented for clarity.

$\mu = 4.5$ and $\sigma = 0.5$ for T and Y types (Morley et al. 2012). This parameter describes the optical thickness of the metal clouds in the brown dwarf atmosphere.

We consider the modeled stellar spectra from Baraffe et al. (2015) at $1200 < T_{\rm eff} < 7000$ K (BT-Settl, CIFIST2011_2015), Morley et al. (2012) at $500 < T_{\rm eff} < 1200$ K and Morley et al. (2014) at $200 < T_{\rm eff} < 500$ K. These are physically-motivated high-resolution spectra from optical to mid-infrared, including absorption by water, methane, ammonia and metal clouds. We extrapolate the templates blueward of 6000 Å with a blackbody spectrum at the corresponding effective temperature if necessary. Cold brown dwarf spectra differ from blackbodies by several orders of magnitudes below 1 μm (Morley et al. 2014), hence we scale the blackbody spectrum to the bluest template point. We assign to each parameter set ($T_{\rm eff}$, $\log g$, $f_{\rm sed}$) the template with the closest parameters. We check that our modeling can reproduce the optical and near-IR magnitudes from the Besançon model output for F to M stars. Emitted spectra are finally scaled according to the stellar radii and heliocentric distances.

3. Methodology

3.1. Programs

In this paper, we consider two accepted JWST observing programs in the Extended Groth Strip (EGS) and the *Hubble* Ultra-Deep Field (HUDF). The existing HST imaging data in the optical and near-infrared are utilized in both fields. We exclusively simulate high-resolution space-based images and neglect ancillary ground-based data.

3.1.1. Cosmic Evolution Early Release Science survey

The Cosmic Evolution Early Release Science (CEERS[8]; P.I.: S. L. Finkelstein) survey is one of the JWST Early Release Science (ERS) programs. CEERS includes multiple imaging (NIRCam, MIRI) and spectroscopic observations over 100 arcmin^2 in the EGS HST legacy field. As shown in Fig. 5, the mosaic

[8] https://jwst.stsci.edu/observing-programs/approved-ers-programs

Table 1. Summary of the JWST imaging data in CEERS and the HUDF – limiting magnitudes.

Name	Area [a]	NIRCam									MIRI	
	[arcmin^2]	*F090W*	*F115W*	*F150W*	*F200W*	*F277W*	*F335M*	*F356W*	*F410M*	*F444W*	*F560W*	*F770W*
CEERS_1	96.8	–	28.7	28.9	29.1	29.2	–	29.2	–	28.7	–	–
CEERS_2	4.6	–	28.7	28.9	29.1	29.2	–	29.2	–	28.7	25.9	25.9
HUDF_1	4.7	29.9	30.3	30.3	30.3	30.6	29.9	30.5	29.9	30.1	–	–
HUDF_2	2.3	29.9	30.3	30.3	30.3	30.6	29.9	30.5	29.9	30.1	28.1	–

Notes. The magnitudes are 5σ point-source limits measured in 0.2″ and 0.6″ diameter apertures for NIRCam and MIRI, respectively. [a]The configurations without MIRI include the area of the configurations with MIRI.

Table 2. Summary of the HST imaging data – limiting magnitudes.

Field	Area [a]	ACS					WFC3			
	[arcmin^2]	*F435W*	*F606W*	*F775W*	*F814W*	*F850LP*	*F105W*	*F125W*	*F140W*	*F160W*
EGS	205	–	28.8	–	28.2	–	–	27.6	26.8 [b]	27.6
XDF	4.7	29.8	30.3	30.3	29.1	29.4	30.1	29.8	29.8	29.8

Notes. The magnitudes are 5σ limits measured in empty circular apertures of diameter 2× the PSF FWHM. [a]This corresponds to the WFC3 surveyed area. [b]The WFC3/*F140W* band in the EGS field is not used in this paper (see Sect. 3.1.1).

pattern consists of ten adjacent and nonoverlapping NIRCam imaging pointings (each NIRCam pointing includes two parallel and separated fields), covering the $1 < \lambda < 5\,\mu$m wavelength range, with four MIRI imaging parallels giving two NIRCam-MIRI overlaps. The estimated 5σ depths are ∼29 mag for NIRCam and ∼26 mag for MIRI (with 32 hours of science integration time). For simplicity, we treat the two distinct observing strategies listed in Table 1. These are the shallowest NIRCam-only and NIRCam-MIRI configurations of the survey, though all the pointings have similar filter choices and exposure times.

The EGS field is supported by the HST/CANDELS multi-wavelength data (Stefanon et al. 2017). We consider the high-resolution HST imaging in the ACS/*F606W*, *F814W* and WFC3/*F125W*, *F160W* bands (Grogin et al. 2011; Koekemoer et al. 2011), as indicated in Table 2. These images reach the 5σ depths of 28.8, 28.2, 27.6 and 27.6 mag, respectively, measured in empty 0.24″, 0.24″, 0.38″ and 0.4″ diameter apertures. We do not use the WFC3/*F140W* imaging from 3D-HST (Brammer et al. 2012; Momcheva et al. 2016) because of its nonuniform layout. In the future, the Ultraviolet Imaging of the CANDELS Fields (UVCANDELS; P. I.: H. Teplitz) will provide deep WFC3/*F275W* and ACS/*F435W* imaging in the EGS field, covering most of the WFC3 footprint and reaching about 27 and 28 mag depths, respectively. These data are not simulated either.

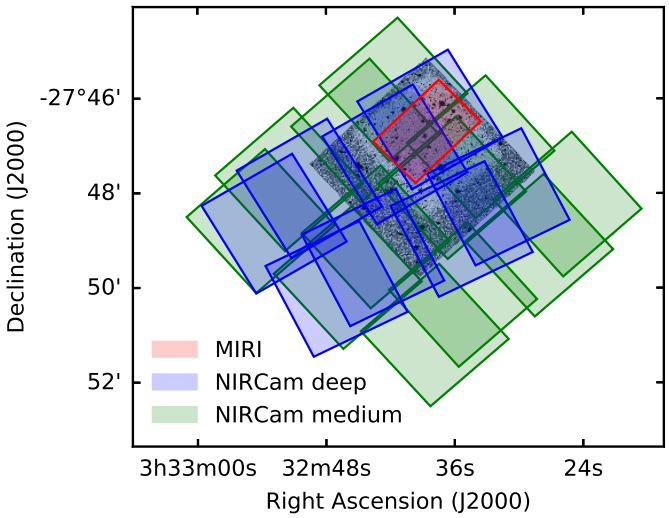

Fig. 6. NIRCam GTO and MIRI GTO layouts in the HUDF. The 4 NIRCam deep pointings are shown in blue, the 6 NIRCam medium pointings in green and the MIRI pointing in red. The ancillary WFC3/IR H_{160}-band coverage in the XDF field is in gray, and the deepest WFC3 region in light gray. The whole XDF field is covered with the deepest ACS data. The dither patterns and the parallel observations are not represented for clarity.

3.1.2. *Hubble* Ultra-Deep Field

Two programs of the Guaranteed Time Observations (GTO) teams are designed to observe the CANDELS GOODS-South field, both of them including deep imaging of the eXtreme Deep Field (XDF). The NIRCam-NIRSpec Galaxy Assembly Survey (P.I.: D. J. Eisenstein) in the GOODS-South and GOODS-North fields includes deep NIRCam preimaging of the HUDF for spectroscopic follow-ups, separated into "Deep" and "Medium" pointings as shown in Fig. 6. This program covers $0.8 < \lambda < 5\,\mu$m with broad-band imaging and two additional medium bands at 3.35 and 4.10 μm. In the GOODS-South field, the "Deep" ("Medium") survey covers 26 (40) arcmin^2 with 174 (42) hours of science time integration, and consists of four (six) NIRCam pointings. In both Deep and Medium pointings (separately), about one third of the area includes overlapping pointings. With NIRCam alone, Williams et al. (2018) predicted several thousands of detected galaxies at $z > 6$ and tens at $z > 10$ at <30 mag (5σ) within the total ∼200 arcmin^2 survey in the GOODS-South and GOODS-North fields. The MIRI HUDF Deep Imaging Survey (P.I.: H. U. Norgaard-Nielsen) consists of MIRI imaging in the *F560W* filter across the 2.3 arcmin^2 of the MIRI field of view. This survey will reach depths of 28.3 mag (4σ) with 49 hours of science integration time for a

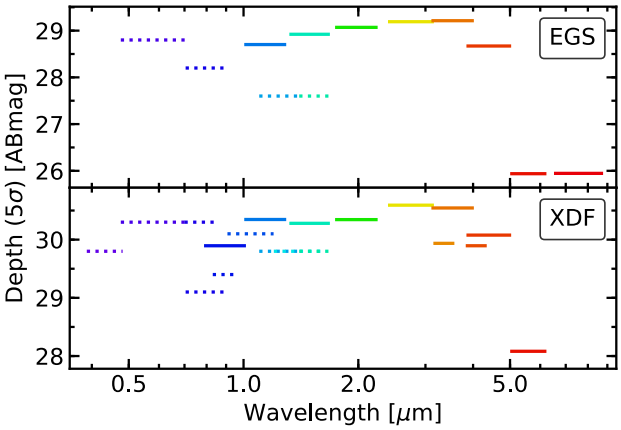

Fig. 7. Limiting magnitudes at 5σ in the simulated data sets in the EGS field (*top*) and the XDF (*bottom*). The list of bands and depths are listed in Tables 1 and 2. The solid lines represent the JWST bands and the dotted lines represent HST bands. The length of each segment is the FWHM of the filter transmission curve.

total of 60 hours. Its layout will be entirely covered by NIRCam imaging.

This field benefits from existing HST/ACS and WFC3/IR imaging, especially in the XDF with the deepest HST imaging ever achieved (Illingworth et al. 2013; Guo et al. 2013). These images cover the optical and near-IR domain $0.38 < \lambda < 1.68\,\mu$m with 9 filters, across 10.8 (4.7) arcmin^2 for the deepest optical (infrared) data. In most filters, the typical 5σ depth reaches 30 mag in the deepest region, measured in empty 0.35″ diameter apertures. We consider two configurations in the deepest WFC3/IR region as described in Table 1, combining the NIRCam Deep and Medium pointings without the respective overlaps, and either with or without MIRI. The bands and depths are listed in Tables 1 and 2, and represented in Fig. 7.

3.2. Mock image simulation

Mock images are generated with SkyMaker (Bertin 2009) from the convolution of point-like and extended sources (with any Sérsic index) with an external point spread function (PSF). Modeled PSF images are created with webbpsf[9] for JWST and TinyTim[10] (Krist et al. 2011) for HST, with an oversampling of five to avoid aliasing effects. The G2V star spectrum from Castelli & Kurucz (2004) is taken as source to generate polychromatic PSFs. In mock HST images, we include an additional jitter (Gaussian blurring) tuned to recover the measured PSF full width at half maximum (FWHM) in real data. The modeled PSF files are multiplied by a radial Fermic-Dirac kernel to limit edge effects around bright sources. Our noise model consists of a single (uncorrelated) Poisson component for photon noise. In real images we expect the noise to be sky-dominated especially for faint sources, we therefore neglect other noise components such as readout noise, inter-pixel capacitance and cosmic rays. Background levels and detection limits can be estimated with the Exposure Time Calculators (ETC) for HST[11] and JWST[12] (Pontoppidan et al. 2016). We tune the background surface brightnesses to reproduce the predicted or measured depths in each band.

[9] https://webbpsf.readthedocs.io/en/stable/
[10] http://www.stsci.edu/software/tinytim/
[11] http://etc.stsci.edu/
[12] https://jwst.etc.stsci.edu/

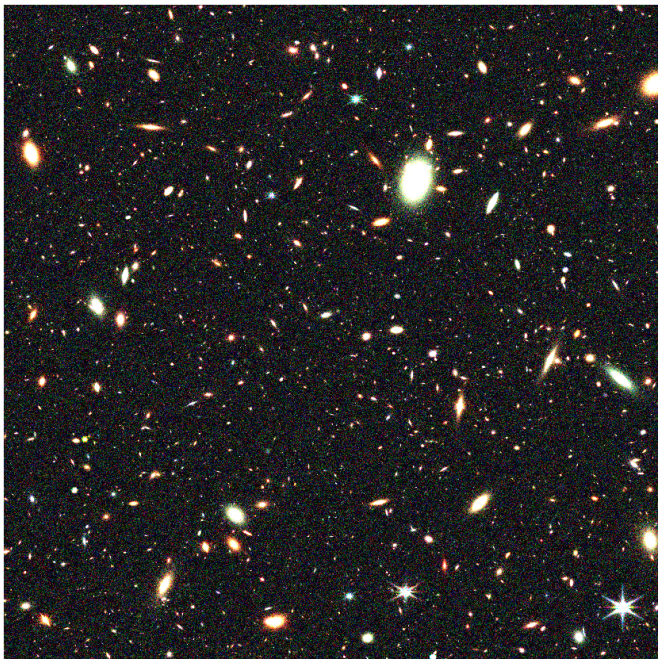

Fig. 8. Simulated composite image in the NIRCam/$F115W$, $F200W$ and $F356W$ bands to a depth of ∼29 mag (5σ), following the CEERS observing strategy. The area is 4.5 arcmin^2 and the resolution 0.031″/pixel for all the images (see Sect. 3.2).

In each of the considered observing strategies, we generate mock images of 11×11 arcmin^2 including all the sources from the mock catalogs. The resulting predictions are then scaled down by numbers to match the area of the planned observations. We generate all images directly at the NIRCam short-wavelength pixel scale of 0.031″/pixel (the smallest among the instruments), avoiding astrometric and resampling issues. Saturation effects are neglected, since the effective saturation limit of stacked small exposure images, as well as the detector nonlinearity, may be difficult to model. Figure 8 shows an example of a simulated composite image in three NIRCam bands. Real images from future JWST surveys may depart from our mock images because of neglected instrumental effects.

3.3. Source extraction

Photometry is measured with SExtractor (Bertin & Arnouts 1996) in the dual-image mode. We successively use the NIRCam/$F115W$, $F150W$ and $F200W$ images as detection image, then combine the extracted catalogs with a 0.2″ matching radius. We use the "hot mode" SExtractor parameters from Galametz et al. (2013) optimized for faint sources, and we checked that we can effectively recover the sources detectable by eye. The redshift distribution of the detected galaxies in the CEERS and the HUDF configurations are represented in Fig. 2. We do not mask bright sources.

Aperture photometry generally provides the less noisy color measurements compared to Kron (Kron 1980) photometry MAG_AUTO (Hildebrandt et al. 2012), however this requires the images to be PSF-matched. We compute the PSF-matching kernels to the HST/$F160W$ band PSF with pypher[13] (Boucaud et al. 2016), from the PSF files used to create the mock images (neglecting PSF reconstruction). Each initial PSF file

[13] https://github.com/aboucaud/pypher

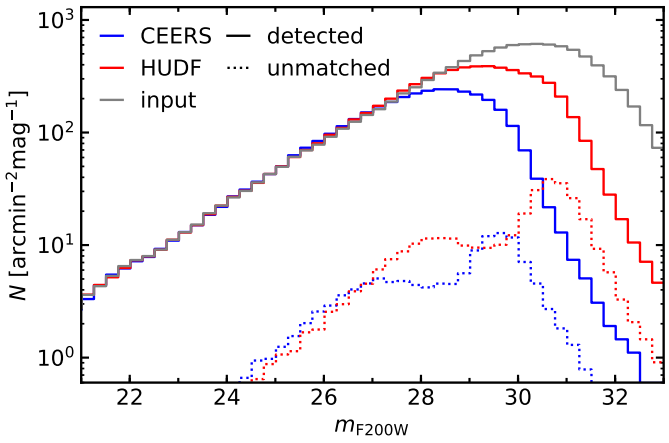

Fig. 9. Detected source number counts versus NIRCam/$F200W$ magnitude. The gray line indicates the input magnitudes of all the mock galaxies. The colored lines illustrate the measured magnitudes of detected sources in our CEERS and HUDF simulations. The solid lines include all the detected sources, the dashed lines represent unmatched sources only.

is resampled to the pixel scale of the target PSF, then the kernel is computed, resampled to the image pixel scale and convoluted with the image (Aniano et al. 2011). Fluxes are measured in 0.5″ diameter apertures (McLure et al. 2013; McLeod et al. 2015, 2016), ensuring at least 70% point-source flux is included in all bands. The PSF FWHM (respectively the 80% encircled energy radius for point source) are 0.145″ (0.28″) for the NIRCam 4.4 μm band and 0.25″ (0.49″) for the MIRI 7.7 μm band.

Following Laigle et al. (2016), we apply corrections to the aperture photometry. SExtractor is known to underestimate flux errors in the case of correlated noise (e.g., Leauthaud et al. 2007), arising from PSF-matching. We therefore apply a band-dependent correction to the measured flux errors, from the ratio of the median flux in empty apertures and the standard deviation of the source flux errors (Bielby et al. 2012). In addition, we scale both fluxes and flux errors with a source-dependent aperture to total correction, computed using MAG_AUTO measurements (Moutard et al. 2016). We exclude truncated photometry and reject objects with negative aperture flux in all bands. Finally, we match the detected object positions with the input source catalog, taking the nearest match within a 0.1″ search radius[14]. Figure 9 illustrates the detected number counts for both of the CEERS and HUDF configurations. The unmatched sources (indicated with dotted lines) present two components, the bright one including artifacts around stars and undeblended sources. We recall that the number of false detections is very sensitive to the noise model.

Galactic foreground extinction remains minimal in extragalactic fields at high galactic latitudes. In practice, it is often corrected by adjusting the image photometric zeropoints. We estimate the zeropoint corrections using the extinction curve of Fitzpatrick (1999) and the Milky Way dust map from Schlafly & Finkbeiner (2011). At the galactic latitude of the EGS field or the HUDF, the correction is at most 0.03 mag in the bluest considered band (ACS/$F435W$), therefore we decide to neglect galactic extinction in both the mock input spectra and the source extraction pipeline.

[14] Sources which are detected beyond this radius are either source pairs or false detections wrongly matched to undetected sources. The probability of the latter event is ∼2%.

3.4. Photometric redshift estimation

To compute photometric redshifts, we perform SED-fitting with LePhare (Arnouts et al. 2002; Ilbert et al. 2006). Following Ilbert et al. (2009), we use the 31 templates including spiral and elliptical galaxies from Polletta et al. (2007) and a set of 12 templates of young blue star-forming galaxies using BC03 stellar population synthesis models. The BC03 templates are extended beyond 3 μm using the Polletta et al. (2007) templates, which include both PAH and hot dust emission from averaged *Spitzer*/IRAC measurements. This set of templates has been extensively tested by the COSMOS collaboration (e.g., Onodera et al. 2012; Laigle et al. 2016) and tested in hydrodynamical simulations (Laigle et al. 2019). We do not include the two templates of elliptical galaxies added in Ilbert et al. (2013) to avoid potential loss of information from degeneracies over the large redshift interval (Chevallard & Charlot 2016). Dust reddening is added as a free parameter ($E(B-V) \leq 0.5$) and the following attenuation laws are considered: Calzetti et al. (2000), Prevot et al. (1984), and two modified Calzetti laws including the bump at 2175 Å (Fitzpatrick & Massa 1986). Nebular emission lines are added following Ilbert et al. (2009). We impose that the absolute magnitudes satisfies $M_V > -24.5$ for CEERS and $M_B > -24$ in the HUDF, based on the LFs at $z < 2$ in Ilbert et al. (2005) and assuming this is still valid at $z > 2$. This SED-fitting prescription (e.g., SFH, attenuation) is distinct from the one used to generate the JAGUAR mock galaxies. This variability may reflect the potential disagreement between the fitted templates and reality, at least to a certain level.

The redshift probability distribution functions (PDFz) are measured in the redshift interval $0 < z < 15$. We perform SED-fitting using fluxes (not magnitudes) and do not use upper limits because this may remove essential information (Mortlock et al. 2012). We add a systematic error of 0.03 mag in quadrature to the extracted fluxes to include the uncertainties in the color-modeling (set of templates, attenuation curves). Photometric redshifts are defined as the median of the PDFz (Ilbert et al. 2013).

Star templates are also fitted to reproduce and quantify potential object misclassification. Similarly to Davidzon et al. (2017), we use the star templates from Bixler et al. (1991), Pickles (1998), Chabrier et al. (2000), the brown dwarfs templates from Baraffe et al. (2015) (see Sect. 2.4) and the BT-Settl grids with Caffau et al. (2010) solar abundances at lower temperatures. These templates partly differ from the set of templates used to generate the mock stars.

We do not attempt to fit AGN templates. SEDs which are AGN-dominated typically present a featureless power-law optical-to-infrared continuum, strong emission lines and Lyman alpha (Lyα) forest absorption especially at high redshifts. The observed emission of galaxies hosting AGNs strongly depends on the contribution of the two components. A large number of hybrid templates would be necessary to correctly characterize them, leading to risks of degeneracies in the SED-fitting procedure (Salvato et al. 2009). In addition, AGNs exhibit variable emission with timescales from minutes to decades. Source variability may be observed from multiple-exposure imaging and dithering in both CEERS and the HUDF, so that AGNs brighter than the detection limit with relatively short timescales should be identifiable.

3.5. Physical parameter estimation

We run LePhare a second time following Ilbert et al. (2015) to determine other physical parameters such as stellar mass (M_*),

star formation rate[15] (SFR) and absolute UV magnitudes ($M_{\rm UV}$). Absolute magnitudes (uncorrected for attenuation) are computed using a top-hat filter of width 100 Å centered at 1500 Å rest-frame (Ilbert et al. 2005). Redshifts are fixed to the photometric redshifts from the first LePhare run. The grid of fitted galaxy templates consists of BC03 models assuming exponential SFHs with $0.1 < \tau < 30$ Gyr, and delayed SFHs ($\tau^{-2}te^{-t/\tau}$) peaking after 1 and 3 Gyr. Two metallicities are considered ($Z_\odot$, $0.5Z_\odot$). We allow $E(B-V) \leq 0.5$ and only include the Calzetti et al. (2000) starburst attenuation curve for simplicity and computational time (Ilbert et al. 2015 included two attenuation curves). Physical parameters are defined as the median of their marginalized probability distribution functions.

4. Physical parameter recovery

4.1. Photometric redshift recovery

The recovery of the photometric redshifts through SED-fitting can first be tested. The quality of the photometric redshifts is assessed with the following statistics (Ilbert et al. 2006): (1) the mean normalized residual $\langle \delta z \rangle$, with the normalized residuals $\delta z = (z_{\rm phot} - z_{\rm true})/(1 + z_{\rm true})$, (2) the normalized median absolute deviation (NMAD) $\sigma_{\rm NMAD} = 1.4826 \times {\rm med}(|\delta z - {\rm med}(\delta z)|)$, and (3) the fraction of catastrophic failures η, for which $|\delta z| > 0.15$.

Figure 10 represents the photometric and true redshifts for all the considered observing strategies, in multiple magnitude intervals. No selection is applied. We observe no systematic bias at $z_{\rm true} < 2$ in any configuration for the bright samples, for which the galaxy continuum redward of the Balmer break is sufficiently well sampled. However, the mean normalized residual becomes negative $\langle \delta z \rangle < -0.1$ at $z_{\rm true} > 2$, even in the brightest magnitude interval. This is probably due to the different attenuation curves in the mock galaxies and in LePhare. The effective, galaxy-wide attenuation curves of the JAGUAR mock galaxies (which employ the two-component attenuation law of Charlot & Fall 2000) are typically grayer (flatter) than the Calzetti et al. (2000) model in the infrared. The bump at 2175 Å in the attenuation curve utilized in LePhare and not JAGUAR may also be an issue.

In the CEERS_1 observing strategies, the number of catastrophic failures is significant even in the brightest magnitude bin. There are several explanations for that. At $z_{\rm true} < 4$, there is a significant number of sources whose redshift is underestimated. Attenuated blue galaxies may be confused with lower redshift unattenuated red galaxies. One of the main reasons for this is the degeneracy between the Lyman and the Balmer breaks, as confirmed from spectroscopic surveys (Le Fèvre et al. 2015). This confusion is enhanced by the lack of optical data in the EGS field, with no deep imaging blueward of HST/$F606W$, so that the Balmer break cannot be correctly identified at low redshift. This is the main reason for the outliers among bright sources. At $z_{\rm true} > 4$ the Lyα break becomes detectable in the HST bands. The number of catastrophic redshift underestimates is therefore reduced, especially for bright sources thanks to the NIRCam bands sampling both the Balmer and Lyα breaks. Strong emission lines may lead to overestimating the continuum, especially for observing strategies which only employ broad-band filters. This can have a significant impact on determining the position of the Balmer break. Quiescent galaxies appear to have a larger dispersion but a smaller outlier fraction than star-forming galaxies.

The two additional MIRI bands at 5.6 and 7.7 μm in the CEERS_2 observing strategy marginally improve the photometric redshift estimates. Both dispersion and outlier rate are larger in the brightest magnitude interval and smaller at fainter magnitudes. At high redshift $z > 4$, the MIRI filters cover the rest-frame near-IR or optical region, therefore sampling the stellar continuum or even the Balmer break. The photometric redshift dispersion is reduced by $\Delta\sigma_{\rm NMAD} = 0.01$ for $4 < z_{\rm true} < 7$ galaxies. Most of the faint NIRCam-detected sources, however, are not detected in MIRI at the depths which will be probed by the CEERS survey. For low-redshift $z < 4$ galaxies, the HST+NIRCam bands impose most of the constraints on photometric redshifts. We still observe fewer catastrophic failures because of Lyman-break misidentification when MIRI data are available, and a systematic bias lowered by 0.05 at $z_{\rm true} = 2$. This comes from a improved sampling of the stellar continuum with MIRI. However, the number of outliers with $z_{\rm true} < 4$ and $z_{\rm phot} > 4$ at $m_{F200W} < 26$ is increased. One of the reasons for more outliers among bright sources with MIRI may be the treatment of dust. The key feature appears to be the observed-frame mid-IR colors. Galaxies with good photometric redshifts mostly present decreasing mid-IR emission with increasing wavelength, whereas outliers often present increasing mid-IR emission. This feature can appear in our mock galaxies from (1) large dust continuum, remaining non negligible even at ~2−3 μm rest-frame because of high dust temperature, or (2) large PAH emission lines at 3.3, 6.2 and 7.7 μm. The infrared luminosities may be overestimated, notably because of the energy balance assumption. In contrast, we are not performing an energy balance in the fitting with LePhare, so that the attenuation and dust emission are disconnected. In addition, the Polletta et al. (2007) templates include dust emission from averaged *Spitzer*/IRAC measurements, so they may not include the mid-IR brightest galaxies. Consequently, LePhare tends to favor high-redshift solutions for low-redshift galaxies with bright and red mid-IR colors. Because of these uncertainties in the mid-IR modeling, one could increase the systematic error added in quadrature to the MIRI photometry. However, this would reduce the additional mid-IR information which is essential to their characterization of high-redshift sources. We therefore do not follow this option.

The main improvements in the HUDF configurations are the deeper HST and NIRCam photometry, leading to the considerable improvements in both the photometric redshift dispersion and outlier rates compared to CEERS. Spectral features such as the Lyman and Balmer breaks can be better captured with the twice more numerous HST bands in the red and near-IR filters. As a consequence, the number of low-redshift galaxies at $z < 3$ with $z_{\rm phot} > 4$ is significantly reduced. Moreover, the additional B_{435} band offers an improved sampling of the Balmer break at $z < 3$ and the Lyman break at $z > 4$. We find that the global outlier rates and photometric redshift dispersion are decreased by about 10% thanks to the addition of the blue band. Furthermore, the two NIRCam medium-bands marginally reduce the redshift outlier rates at $z > 6$ mostly. With the additional MIRI/$F560W$ band in HUDF_2, we do not observe any improvement in the global photometric redshift dispersion or outlier rate. In contrast, both of them are improved at high redshift and especially at $z > 10$, where MIRI provides the only information redward of the Balmer break.

Source blending may also lead to catastrophic photometric redshifts, because of contaminated photometry and incorrect colors. The photometric redshifts of blended high-redshift

[15] In LePhare, the measured SFR is instantaneous, whereas in JAGUAR it is averaged over the past 100 Myr. For exponential and delayed SFH with $\tau > 1$ Gyr, the difference between these two SFR definitions is no more than 5% (0.02 dex) at $z > 0.1$.

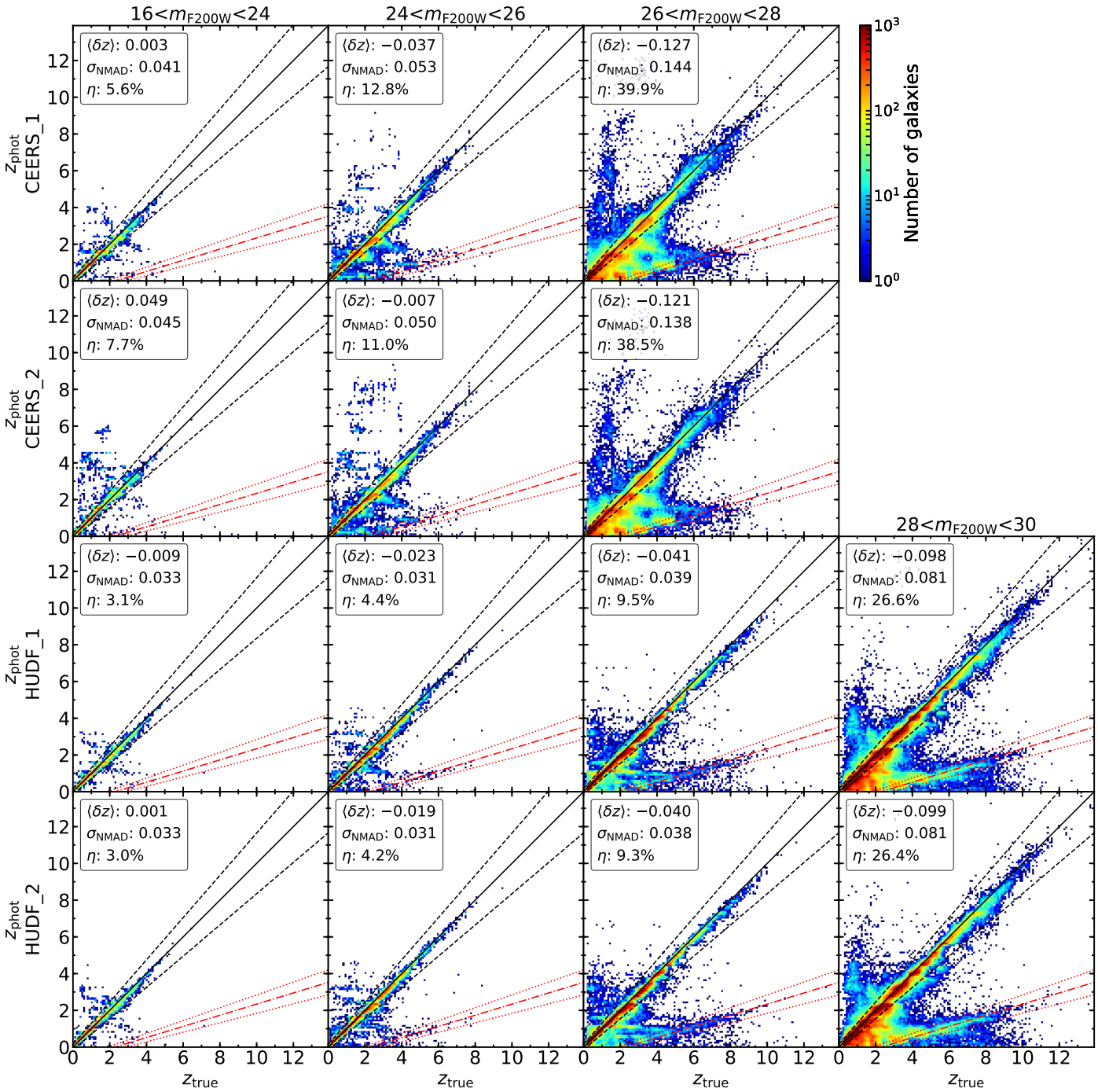

Fig. 10. Comparison between photometric and true redshifts, "true" redshifts being in this case the simulated redshifts as described in Sect. 2.1. The rows correspond to CEERS_1, CEERS_2, HUDF_1 and HUDF_2 observing strategies from top to bottom, and the columns represent observed NIRCam/$F200W$ magnitude intervals. Color indicates the number of sources. The mean normalized residual, the normalized median absolute deviation and the catastrophic error fraction η for all the detected sources are indicated. The solid black line shows the 1:1 relation and the dashed black lines the $\pm 0.15(1 + z_{true})$ threshold used to compute η. The degeneracy between the Balmer 4000 Å break and the Lyα 1216 Å break is identified by the dotted-dashed red line, with 15% errors in dotted red lines.

galaxies tend to be mostly underestimated, which is coherent with blended source pairs or groups which are most likely to contain at least one galaxy at $z \sim 1$–2. Figure 11 illustrates the fraction of detected sources with the SExtractor flags indicating blending (flag 2), contaminated photometry (flag 1), both (flag 3) or none (flag 0). The solid lines represent the photometric redshift outliers and dotted lines the whole detected sample. The number of flagged sources mostly decreases with increasing magnitudes, because faint sources typically need to be isolated to be detected, therefore unflagged. In contrast, faint sources with bright neighbors may remain undetected. Flagged objects represent a large portion of the detected sources, about 40% (65%) at $m_{F200W} < 26$ in CEERS (HUDF), which increases with the depth of the survey. We observe a significantly increased fraction of contaminated sources (flags 1+3) in the z_{phot} outliers, about twice as much as in the entire sample at $m_{F200W} < 26$

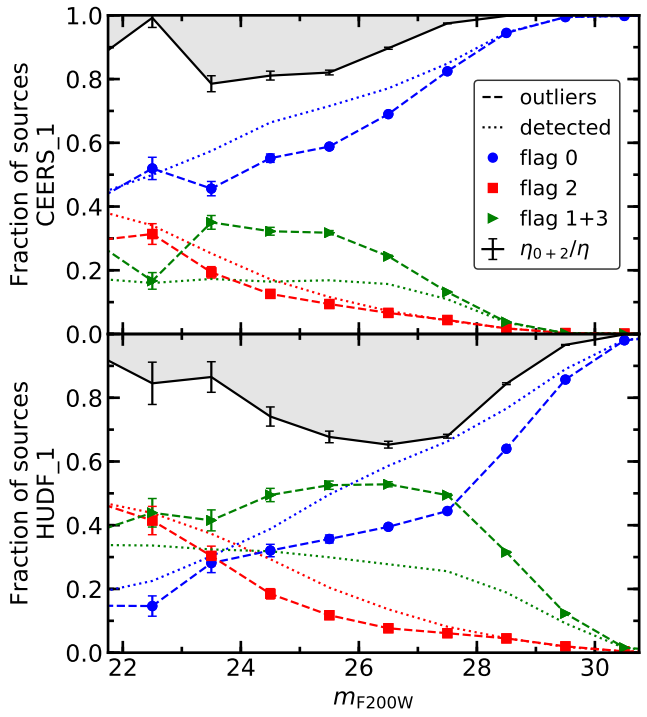

Fig. 11. Fraction of detected sources per SExtractor flags indicating blending (flag 2), contaminated photometry (flag 1), both (flag 3) or none (flag 0) versus observed NIRCam/$F200W$ magnitude. The rows correspond to the CEERS_1 (*top*) and HUDF_1 (*bottom*) observing strategies. The dashed lines indicate the flag fractions among photometric redshift outliers (summing to one), as defined in Sect. 4.1. The dotted lines represent the flag fractions in the whole detected sample (summing to one). The solid black line shows the ratio between the photometric redshift outlier rates η_{0+2} assuming all the sources have uncontaminated photometry (flags 0 or 2), and the standard outlier rates. The error bars are propagated Poisson errors.

in CEERS and at $m_{F200W} < 28$ in the HUDF. In the hypothetical case where all the detected sources had uncontaminated photometry (flags 0+2), the photometric redshift outlier rates η would be corrected by the indicated ratio η_{0+2}/η. This ratio reaches 80% at $24 < m_{F200W} < 26$ in CEERS, meaning that the outlier rate would decrease from 12% to 10% in this magnitude bin. Similarly in the HUDF, the outlier rate at $26 < m_{F200W} < 28$ would decrease from 9% to 6%. We observe no significant wavelength dependence of these values. These results indicate that source blending will definitely be an issue with deep JWST imaging.

4.2. Stellar mass recovery

The comparison between input stellar masses and those measured through SED-fitting is illustrated in Fig. 12 for the CEERS_1 and HUDF_1 observing strategies. The redshift intervals are centered at $z = 1, 2, 3, \ldots$ with a width of $\Delta z = 1$. The measured stellar masses agree well with the input ones.

In the CEERS_1 configuration, the stellar mass dispersion is below 0.25 dex at $\log(M_*/M_\odot) > 9$. Removing the photometric redshift outliers lowers the dispersion at $\log(M_*/M_\odot) > 9$ to 0.2 dex, and significantly reduces the number of catastrophic stellar mass estimates at $\log(M_*/M_\odot) < 8.5$. These low-mass objects are typically fainter and have noisier colors. The overall dispersion increases as the stellar mass decreases, from 0.25 dex to 0.5 dex for $\log(M_*/M_\odot)$ of 9 and 8, respectively, and for galaxies with good photometric redshifts, from 0.2 dex to up to 0.35 dex. Most of the outliers at $z < 4$ have stellar masses which are overestimated because of overestimated redshifts. The remaining cases are galaxies with nearby sources, boosting their aperture fluxes and affecting their colors. For blended galaxies with nevertheless correct colors and photometric redshifts, the total fluxes may not be correctly recovered through the deblending procedure despite the aperture-to-total flux correction. In the HUDF_1 strategy, both the number of outliers and the dispersion are smaller for low-mass galaxies, remaining below 0.45 dex up to $\log(M_*/M_\odot) = 7$ and even below 0.35 dex when discarding catastrophic photometric redshifts. Stellar masses are not significantly affected by redshift outliers above $\log(M_*/M_\odot) > 8$. The dispersion remains below 0.2 dex at $\log(M_*/M_\odot) > 9$ at all redshifts. These results mainly reflect the improvements in the photometric redshifts from deeper HST and NIRCam imaging and with the additional HST blue band. Moreover, the near-IR medium-band photometry enable the emission lines and the galaxy continuum to be better separated. The systematic overestimation and the dispersion at $z \geq 6$ are lowered by 0.05 dex only thanks to the medium bands, which are located close to the Balmer break in the rest-frame.

The median stellar mass lies between ±0.2 dex around the input value between $8 < \log(M_*/M_\odot) < 10$ at all redshifts and in both configurations, and is generally underestimated at $z \leq 3$ and overestimated at $z > 4$. These observations may again come from the steepness at $\lambda > 1\,\mu$m of the attenuation curves used in input and in the SED-fitting (see Sect. 4.1). More attenuation may hide more low-mass stars and therefore result in underestimated mass. At $\log(M_*/M_\odot) < 8$, the stellar mass estimates are systematically overestimated for galaxies with correct redshifts. This bias increases with redshift, and reaches at most 0.8 dex at $\log(M_*/M_\odot) = 7$. At $\log(M_*/M_\odot) > 10$ and $z < 6$, stellar masses are systematically underestimated by 0.15 − 0.2 dex. Massive galaxies are typically the most attenuated, and these galaxies effectively have large input attenuation $\hat{\tau}_V > 0.1$. The percentage of $\log(M_*/M_\odot) > 10$ galaxies with $\hat{\tau}_V > 1$ is 57%, and reaches 70% for the subset where mass is underestimated by at least 0.2 dex. Strong attenuation $E(B-V) > 0.5$ ($A_V > 2$) are not allowed in our LePhare configuration to avoid additional degeneracies between templates. The underestimated attenuation in SED-fitting may lead to underestimated stellar mass. In contrast, galaxies at $\log(M_*/M_\odot) \sim 9$ with $\hat{\tau}_V > 0.2$ have overestimated stellar masses, by 0.1, 0.2, 0.3 dex at $z = 4, 5, 6$ in both CEERS and the HUDF.

Quiescent galaxies have underestimated stellar masses in all the observing strategies, by 0.15 dex at $\log(M_*/M_\odot) > 9$ and by 0.5 dex below. These numbers are not reduced when removing photometric redshift outliers. High-mass quiescent galaxies typically have large metallicity, however observational constraints on the metallicity of low-mass galaxies are lacking. In JAGUAR, low-mass ($\log(M_*/M_\odot) < 8.7 + 0.4z$) quiescent galaxies are assigned random uniform metallicities between $-2.2 < \log(Z/Z_\odot) < 0.24$. The recovered stellar masses of $\log(M_*/M_\odot) < 9$ quiescent galaxies with $\log(Z/Z_\odot) < -0.5$ (> -0.5) are underestimated by up to 0.7 (0.4) dex. This dramatic underestimation of stellar mass for low-mass quiescent galaxies may come from the quiescent galaxy templates in LePhare which do not span the parameter space of the mock galaxies. In particular, only two metallicities ($\log(Z/Z_\odot) = 0, -0.3$) are allowed in the LePhare configuration to avoid degeneracies between templates. In addition, dust attenuation was neglected for low-mass quiescent galaxies in JAGUAR, which may also explain this systematic bias at low masses.

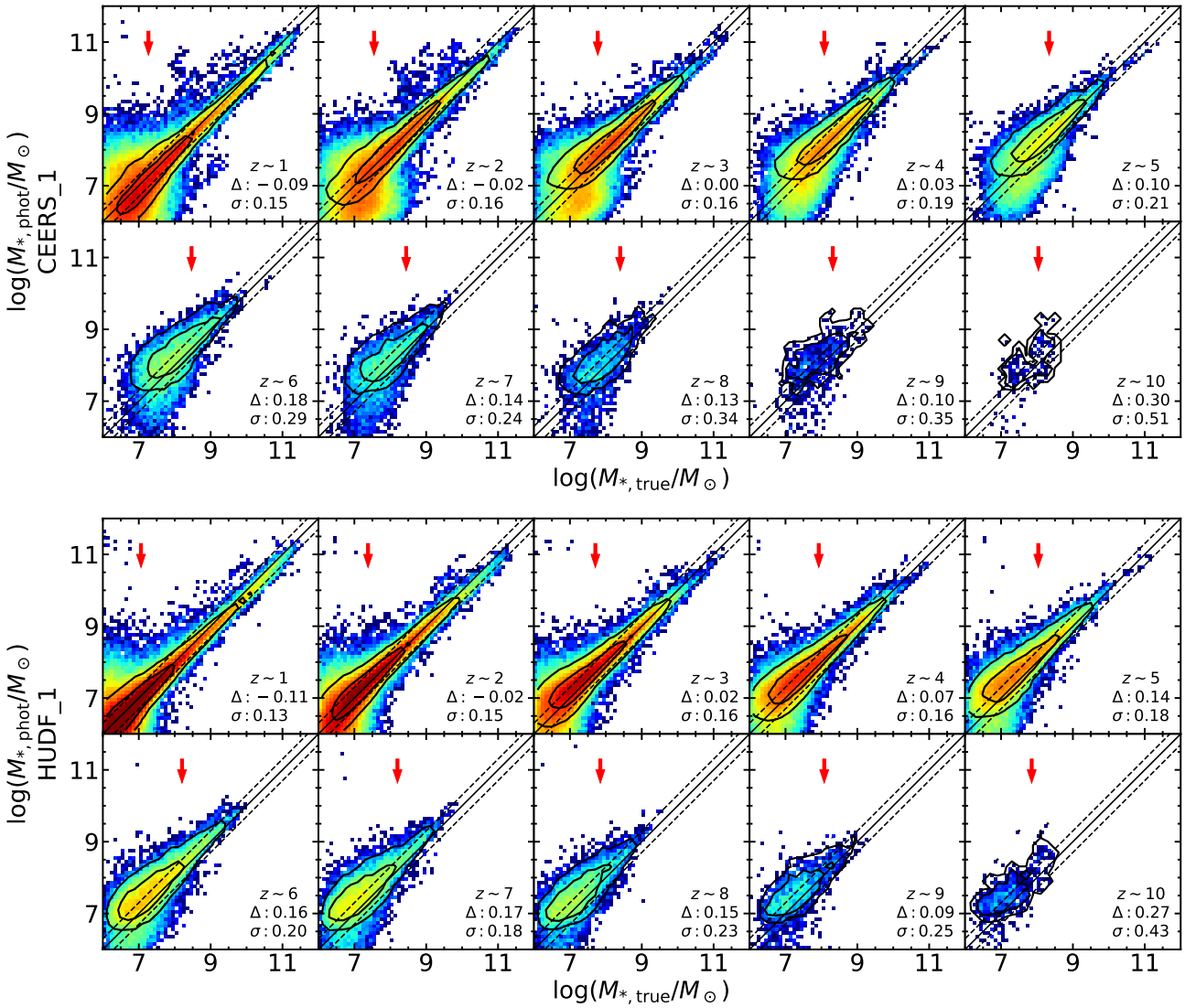

Fig. 12. Comparison between measured and input stellar masses, for the CEERS_1 and HUDF_1 observing strategies (*top and bottom* figures, respectively). Each panel represents an input redshift interval centered at $z_{\rm true} = 1, 2, 3, \ldots$ of width $\Delta z = 1$. Color indicates the number of sources in the whole sample, with the same color scale as in Fig. 10. The thick black contours represent the distribution of the sources with correct photometric redshift $|z_{\rm phot} - z_{\rm true}| < 0.15 z_{\rm true}$, including 68% and 95% of these sources, respectively. The median shift (Δ) and the NMAD (σ) for the sources with correct $z_{\rm phot}$ and $8 < \log(M_{*,\rm true}/M_\odot) < 10$ are indicated (in dex). The solid line shows the 1:1 relation and the dashed lines ± 0.3 dex. The red arrows indicate the detected 90% stellar mass completeness.

The CEERS_2 strategy presents results equivalent to CEERS_1. This is not surprising because of the galaxy continuum already well sampled with NIRCam. At very high redshift $z > 10$ where NIRCam does not sample redward of the Balmer break, the photometric redshift estimates still rely on NIRCam, and the shallow MIRI imaging does not significantly improve the stellar mass estimates. With HUDF_2 however, the MIRI data are deep enough to slightly improve stellar masses at $z \geq 9$, with the scatter and systematic bias lowered by about 0.05 dex. This essentially comes from the improvement of photometric redshifts.

4.3. Star formation rate recovery

Figure 13 illustrates the galaxy star formation rate recovery in the CEERS_1 and HUDF_1 observing strategies. The results with the CEERS_2 and HUDF_2 configurations, respectively, are strictly similar.

The measured SFRs remain in correct agreement with the input values, however less precise than stellar mass estimates. We note that the SFR estimates may behave well because the assumed SFH in LePhare and in JAGUAR are similarly simple, meaning smooth exponential or delayed SFH. The low precision is primarily due to the degeneracy between SFR and dust attenuation, which affects the rest-frame UV, where the emission is dominated by hot, young stars. In an analogous work, Laigle et al. (2019) showed that with a similar LePhare configuration, attenuation is the main source of systematic uncertainties and dispersion in the SFR recovery. In addition, the missing nebular continuum emission in LePhare may also be an issue. For galaxies with good photometric redshifts, the SFR dispersion is 0.3 dex for the CEERS survey and 0.35 dex for HUDF, and remains stable over redshift and input SFR. In the HUDF, however, the recovered SFR distributions are skewed toward large SFRs at $z \geq 3$. This surely comes from the more difficult match between the input and the fitted galaxy templates,

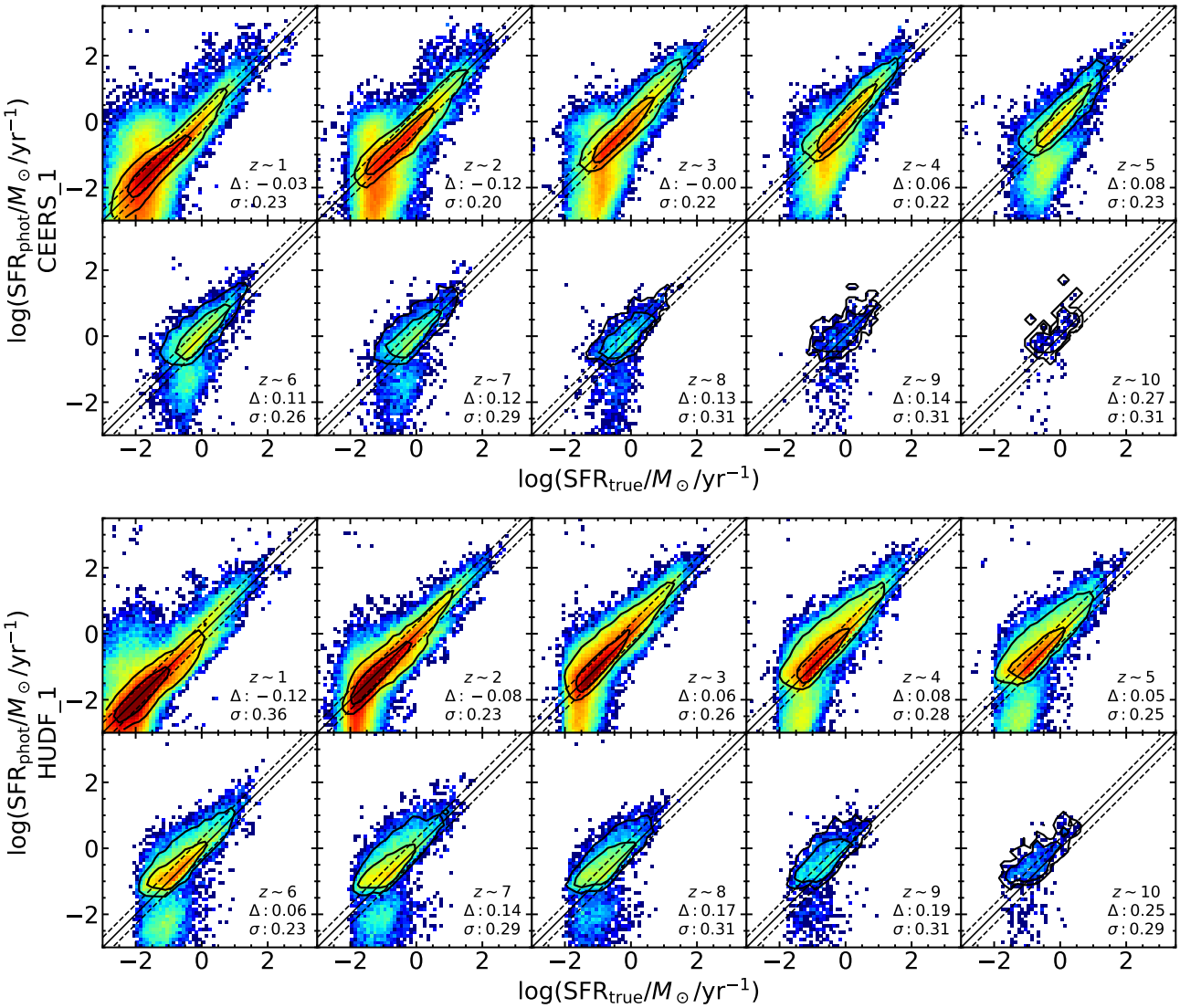

Fig. 13. Same as Fig. 12, but for star formation rate. The median shift (Δ) and the NMAD (σ) for the sources with correct $z_{\rm phot}$ and $-1 < \log(\mathrm{SFR_{true}}/M_\odot/\mathrm{yr}^{-1}) < 1$ are indicated (in dex). The solid line shows the 1:1 relation and the dashed lines ±0.3 dex.

because of the increased depth in the HUDF and, at low redshift, the HST *B*-band, giving stronger constraints on the SFR tracers.

We observe that the median shift at $\mathrm{SFR_{true}} > 1\,M_\odot\,\mathrm{yr}^{-1}$ is bounded by ±0.2 dex at all redshifts in all the observing strategies. In particular, the most star-forming galaxies at $z < 6$ with $\mathrm{SFR_{true}} > 10\,M_\odot\,\mathrm{yr}^{-1}$ have systematically overestimated SFR estimates by 0.15 dex. This may come from the difference of attenuation curves assumed in JAGUAR and in LePhare. Additionally, most of the outliers at $\mathrm{SFR_{true}} > 1\,M_\odot\,\mathrm{yr}^{-1}$ have overestimated SFR estimates, similarly to the stellar mass outliers. Among galaxies with correct photometric redshifts, the systematic bias increases with decreasing $\mathrm{SFR_{true}}$ and as redshift increases, reaching 0.5 dex at $0.1\,M_\odot\,\mathrm{yr}^{-1}$ in CEERS and 0.4 dex in the HUDF. The large number of catastrophic failures at $\mathrm{SFR_{true}} < 1\,M_\odot\,\mathrm{yr}^{-1}$ at all redshifts comes from redshift misestimation. This feature appears in all the observing strategies, and its importance is only slightly reduced in the HUDF compared to CEERS. In our methodology to estimate galaxy physical parameters, imposing underestimated redshifts in the second SED-fitting run gives underestimated SFRs and vice versa. This would be a priori unknown in real surveys, so it shows the importance of simulations to make the necessary corrections.

4.4. Absolute UV magnitude recovery

Figure 14 illustrates the recovery of absolute UV magnitudes. There are features in common with the stellar mass and SFR measurements, such as outliers with mostly overestimated luminosities, and the dispersion from catastrophic photometric redshifts. In the CEERS configurations, the dispersion increases from 0.2 mag at $M_{\rm UV} = -20$ to 0.3 mag at $M_{\rm UV} = -18$ for $z \leq 3$ galaxies. At higher redshift, the distributions are typically 0.1 mag broader. The UV luminosities are overestimated by 0.15 mag at $z \sim 1$ and underestimated by at most 0.2 mag at $z \geq 2$ for sources with $M_{\rm UV} > -18$ and good photometric redshifts. In comparison, in the HUDF, the dispersion at $M_{\rm UV} < -18$ remains below 0.2 mag at all redshifts for sources with correct photometric redshifts, and below 0.25 mag (0.4) at $M_{\rm UV} < -17$ and $z \leq 3$ (≥ 3). The magnitudes are systematically overestimated by ∼0.1 mag at $M_{\rm UV} > -18$. For low-redshift $z < 2$ galaxies, the improvements in the HUDF are driven by the additional B_{435}-band photometry and the smaller *K*-corrections required to compute the absolute UV magnitudes. In the JAGUAR galaxies, the birth cloud component of the dust attenuation may strongly affect the rest-frame UV emission. LePhare may underestimate

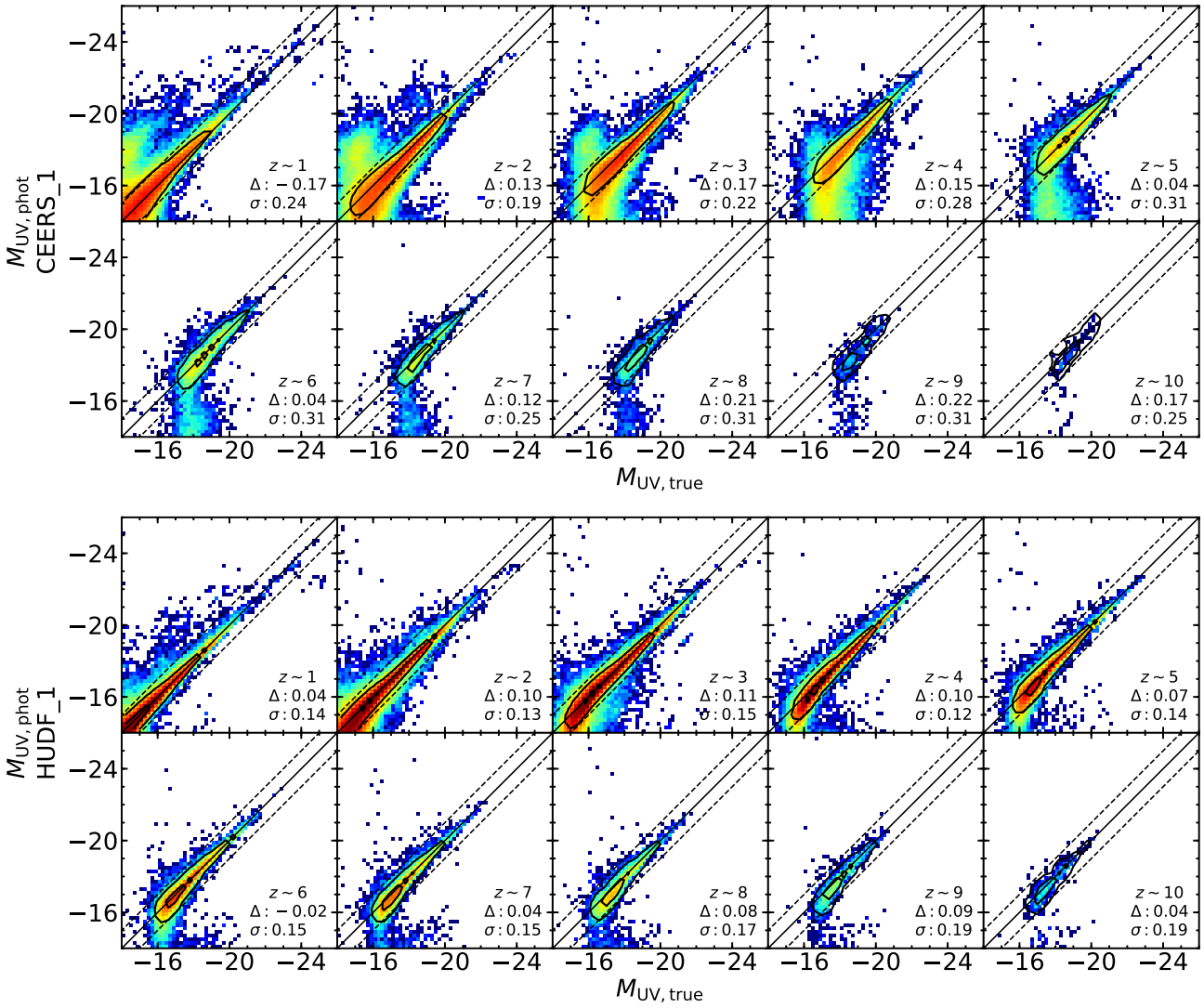

Fig. 14. Same as Fig. 12, but for absolute UV magnitude. The median shift (Δ) and the NMAD (σ) for the sources with correct $z_{\rm phot}$ and $-20 < M_{\rm UV} < -18$ are indicated (in mag). The solid line shows the 1:1 relation and the dashed lines ±1 mag.

the attenuation especially at this wavelength, leading to underestimated UV luminosities. The NIRCam medium bands decrease the systematic bias and dispersion by about 0.03 mag at $z \geq 6$.

4.5. Comparison with previous works

Bisigello et al. (2016, 2017, 2019) investigated the recovery of the galaxy photometric redshifts and physical parameters with JWST broad-band imaging. The authors considered multiple galaxy samples, observed galaxies at $z < 7$ and simulated spectra constructed from BC03 and Zackrisson et al. (2011) population synthesis models at $z > 7$. All the combinations of a few discrete physical parameters were used to build the high-redshift galaxy samples. As a consequence, the distribution of these parameters among the real galaxy population at the given redshift was not respected. In addition, the source samples did not reproduce the redshift distribution of a flux-limited galaxy population, meaning that the contamination from foreground low-redshift galaxies into the high-redshift samples could not be estimated. The galaxy physical properties were then determined through SED-fitting with LePhare using the same galaxy templates as for the input spectra. Stellar and brown dwarf templates were not fitted, meaning that the nature of the sources was assumed to be known a priori. Bisigello et al. (2016) already showed that HST short-wavelength optical data could significantly reduce the photometric redshift dispersion and outlier rate. Bisigello et al. (2017) notably investigated the stellar mass recovery for $7 < z < 10$ galaxies with the eight NIRCam broad-bands and MIRI imaging. The recovered precision on stellar masses were similar to our results, as well as the systematic overestimation attributed to emission lines.

Kemp et al. (2019) analyzed the redshift and stellar mass recovery with JWST and HST imaging. The authors notably used the Empirical Galaxy Generator (EGG, Schreiber et al. 2017) to generate a complete magnitude-limited sample of $0 < z < 15$ and $5 < \log(M_*/M_\odot) < 12$ galaxies over 1.2 deg^2. This catalog included individual spectra with no emission lines, nonetheless the case of emission lines was treated with another sample of mock galaxies. The authors introduced and investigated two observing strategies including eight NIRCam bands with MIRI/$F770W$ parallels, and HST/V_{606} and i_{814} bands as ancillary data. These configurations are similar to the CEERS program, with similar choices of filters, exposure times and depths in the deepest regions. We come to the similar

conclusions about MIRI, namely that its addition leads to an improvement in the photometric redshift recovery at $4 < z < 7$, though most of the constraints are coming from NIRCam and HST. The authors quantified the gain from additional deep HST/B_{435} imaging, which was revealed to be more important than whether MIRI imaging was available.

5. Source selection

In this section, we investigate the selection of high-redshift galaxies and the rejection of contaminants, then we give predictions about the number counts and the recovery of the galaxy luminosity function. The impact of the selection on the galaxy samples is assessed using galaxy completeness and purity. We define completeness as the fraction of input galaxies, in a given magnitude m and redshift z bin, which are selected and assigned to the correct redshift interval. Likewise, purity is the fraction of selected sources, in an observed magnitude $m^{\rm obs}$ and redshift z bin, which are high-redshift galaxies in this redshift interval. We define the redshift intervals $[z_i \pm \Delta z/2]$ with a width of $\Delta z = 1$ and centered at $z_i = 1, 2, 3, \ldots$ Let $N_{\rm input}$ be the number of input galaxies, $N_{\rm selected}$ the number of selected sources assigned to a redshift bin, and $N_{\rm correct}$ the number of selected galaxies which are assigned to the correct redshift interval. $N_{\rm selected}$ may include false detections. Completeness C and purity P can be written as:

$$C(m, z) = \frac{N_{\rm correct}(m, z)}{N_{\rm input}(m, z)}, \quad (1)$$

$$P(m^{\rm obs}, z) = \frac{N_{\rm correct}(m^{\rm obs}, z)}{N_{\rm selected}(m^{\rm obs}, z)}. \quad (2)$$

The number of detected objects depends on the observation and the source extraction, setting the maximum number of sources which can be recovered. There is then a trade-off between completeness, purity and sample size: no selection will give maximum completeness (maximum sample size) and likely minimum purity, whereas stringent selections will lower completeness (lowering sample size) and likely higher purity.

5.1. Star rejection

We first investigate the rejection of stellar objects contaminating the high-redshift galaxy samples. Commonly used criteria rely either on magnitude, colors, shape (or surface brightness) and the quality of the SED-fitting. We consider the following list of standard star rejection criteria, and we investigate the individual impact of each of them:

(i) $S < k$, with $k = 0.95, 0.9$
(ii) $(\mu_{\rm max} > 0.95 m_{F115W} - 1.9) \vee (\mu_{\rm max} > k)$ in CEERS, $(\mu_{\rm max} > 0.95 m_{F115W} - 2.2) \vee (\mu_{\rm max} > k)$ in the HUDF, with $k = 24, 25$
(iii) $(m_{F115W} - m_{F356W} > 0.7(m_{F606W} - m_{F115W}) - 1.55) \vee (m_{F115W} - m_{F356W} > 0.1(m_{F606W} - m_{F115W}) - 0.95)$ if $S/N_{F606W} > 2$, $(m_{F150W} - m_{F200W} > 0.25(m_{F200W} - m_{F444W}) - 0.75) \vee (m_{F150W} - m_{F200W} > 0)$ otherwise
(iv) $\chi^2_{\rm star} > k$, with $k = 0.5\nu, \nu$
(v) $\chi^2_{\rm gal} - \chi^2_{\rm star} < k$, with $k = \nu, 0$

with S defined as the source stellarity index, $\mu_{\rm max}$ is the maximum surface brightness, $\chi^2_{\rm star}$ and $\chi^2_{\rm gal}$ are the mean squared error from the SED-fitting of stellar and galaxy templates, respectively, ν is the number of degrees of freedom in the fitting (set to the number of bands minus three). The thresholds k define a soft (first value) and a stringent (second value) version for some selections. The symbol $\vee$ represents the logical OR.

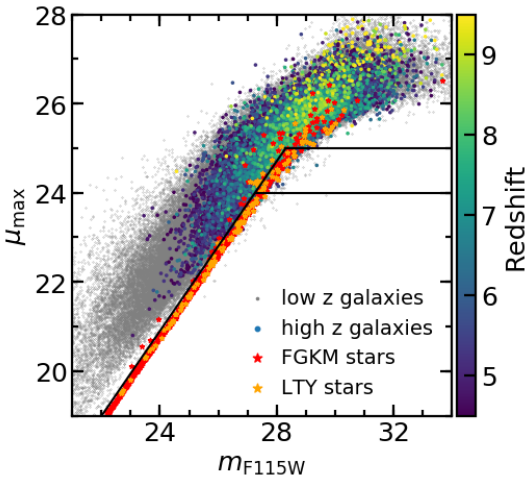

Fig. 15. Maximum surface brightness-magnitude selection criteria to remove stellar objects. Each marker represents the measured colors of a detected source in the CEERS_1 observing strategy. The colored points are high-redshift galaxies and the gray points indicate $z_{\rm true} < 4.5$ galaxies. The red and orange stars represent FGKM and LTY stars, respectively.

The first criterion (i) is based on the stellarity index S measured with SExtractor in the NIRCam/$F200W$ detection image. This is the posterior probability of a detected object to be a point-source (0 for extended source, 1 for point-source), according to its surface brightness profile. With high resolution imaging, brown dwarfs may be separated from resolved galaxies based on size (Tilvi et al. 2013). However, distant galaxies commonly appear point-like (e.g., bright star-forming blobs, faint galaxy hosting a bright AGN) and should not be discarded. The impact of this selection on galaxy completeness therefore depends on the morphology of the galaxies. This will need to be further investigated with simulations of more realistic galaxy light distributions. Similarly, stars tend to occupy a tight locus in the size (or surface brightness) – magnitude plane (Leauthaud et al. 2007). We construct the selection (ii) in the maximum surface brightness $\mu_{\rm max} - m_{F115W}$ plane as represented in Fig. 15. The parameter $\mu_{\rm max}$ is the surface brightness [mag arcsec^{-2}] of the brightest pixel belonging to the source, above the estimated background. The NIRCam/$F115W$ band is well adapted since stars mainly become fainter in redder bands and the emission of MLTY dwarfs drops in bluer bands. We then make use of the color-color selections (iii) following Davidzon et al. (2017). The adopted color diagrams are (HST/$F606W$–NIRCam/$F115W$) vs. (NIRCam/$F115W$–$F356W$) for objects detected at 2σ in the HST/$F606W$ band, and (NIRCam/$F150W$–$F200W$) vs. (NIRCam/$F200W$–$F444W$) for the other sources (Fig. 16). Finally, we consider selections based on the SED-fitting results, either (iv) the absolute quality of the stellar fit (Bowler et al. 2015), or (v) the relative quality of the stellar fit with respect to the galaxy fit (Ilbert et al. 2009).

Figure 17 illustrates the photometric redshift distribution of stellar objects in the CEERS_1 configuration, and the number of remaining stars after each rejection criterion is individually applied. In addition, the resulting differences of purity and completeness for the galaxy samples are indicated, for each redshift interval and integrated over magnitude. The purpose is to remove as many stellar contaminants as possible while maintaining a

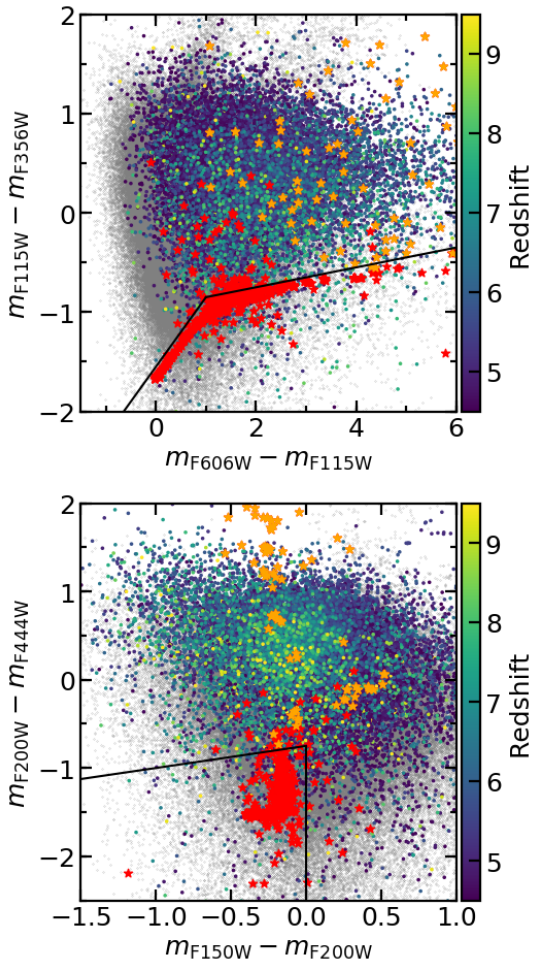

Fig. 16. Color-color selection criteria to remove stellar objects. Each marker represents the measured colors of a detected source in the CEERS_1 observing strategy. The colored points are high-redshift galaxies and the gray points indicate $z_{true} < 4.5$ galaxies. The red and orange stars represent FGKM and LTY stars, respectively. Only sources detected at 2σ in the two reddest bands (in each panel) are indicated.

high galaxy completeness, and any gain in galaxy purity is an additional advantage in terms of statistics of the recovered galaxy population. As expected, the stellarity index cuts (i) manage to efficiently reject stars, about 65% (80%) for $S < 0.95$ (0.9), however lowering galaxy completeness of about 20% (60%) at all redshifts $z > 4$. In contrast, the surface brightness-magnitude selections (ii) remove a similar number of stars and maintain a high completeness and purity. Again, these impact on galaxy completeness depends on the assumed galaxy morphologies. The color-color cuts (iii) have a marginal effect on brown dwarfs with $z_{phot} > 5$, whereas most of the stellar objects with $z_{phot} < 2$ are effectively removed. Neither galaxy completeness nor purity are much affected. The optical and near-IR colors of cold brown dwarfs appear not to occupy the same stellar locus as hotter stars, and removing them in the color-color space would discard many galaxies at the same time. Finally, the criteria based on the absolute quality of the stellar fit (iv) only reject about 30% of the stars, though slightly modifying completeness and purity. In contrast, the selection with the difference of chi squares (v) removes 95% of the stars at all photometric redshifts, maintaining a solid completeness only lowered by 2% and even removing extra contaminants. We find similar results for the other observing strategies.

From these results, we can conclude that the combination of both the soft (ii) and the soft (v) criteria is the most efficient way of removing stars from the high-redshift galaxy candidates. This is used in the next sections. The remaining stellar contaminants in the $z > 4$ galaxy samples decrease from 0.26 ± 0.02 to 0.010 ± 0.004 arcmin^{-2} for CEERS_1 and from 0.18 ± 0.02 to 0.004 ± 0.003 arcmin^{-2} for HUDF_1. The differences between CEERS and the HUDF include the input density of stars at the respective sky coordinates, the depth and wavelength coverage of the observations. The addition of MIRI imaging improves the photometric redshifts of stars, with detected densities of 0.24 ± 0.02 arcmin^{-2} for CEERS_2 and 0.14 ± 0.02 arcmin^{-2} for HUDF_2. These lower values come from the mid-IR colors of stars which are less comparable to galaxy colors than in the near-IR. It should be mentioned that these selection criteria are specifically constructed to reject stars, nevertheless they are not the only criteria to have this effect. Color-color selections designed to select high-redshift galaxies based on their Lyman break may result in extra stellar rejection, whereas SED-fitting-based selections mainly rely on the types of criteria considered above. The final stellar rejection therefore depends on the entire set of selection criteria.

5.2. Galaxy selection at $z > 5$

In this section, we explore multiple procedures to select high-redshift galaxies and estimate the respective impact on galaxy completeness and purity. We use an alternative, more permissive definition of purity in this section only. A selected source, assigned to the redshift interval centered at z_i, is considered as a contaminant if $z_{true} < z_i - 1$. Hence only low-redshift sources are considered as contaminants. This avoids classifying, for example, $z \sim 6.4$ galaxies which are scattering into our $z \sim 7$ selection as contaminants. We do not treat the specific case of faint Lyman alpha emitters (LAE), typically presenting a strong emission in only one or two bands. The redshift of these galaxies cannot be well constrained without narrow-band imaging or spectroscopy (Dunlop et al. 2013), which are not available here. In addition, we do not include criteria based on visual inspection. This technique may be used to discard sources based on shape or colors to consolidate purity, however with real images the resulting galaxy completeness becomes hard to estimate.

We consider three sets of selection criteria summarized in Table 3. These are based on Bouwens et al. (2015), Bowler et al. (2015) and Finkelstein et al. (2015), adapted to the present set of photometric bands and generalized to multiple redshift intervals. We do not include magnitude cuts. The criteria for the EGS field in Bouwens et al. (2015) rely on initial color-color preselections, then on photometric redshifts confirmation. Because of the lack of optical data, and medium or narrow band imaging, it is not possible to select galaxies with color criteria only. The Lyman break galaxies (LBG) color-color selections are represented in Fig. 18. The location of the Lyman alpha break relies on the V_{606} and i_{814} HST bands at $z < 8$. The galaxy colors redward of the break are quantified with NIRCam bands to take advantage of the increased depths compared to the WFC3 bands. Lower-redshift contaminants are expected to be excluded by imposing no detections ($S/N < 2$) blueward of the break. However, this fact is strictly valid at $z > 6$ where the IGM transmission is extremely low. The resulting high-redshift samples may be biased toward young UV bright sources and miss a significant fraction of the galaxies (Hughes et al. 1998; Le Fèvre et al. 2015; Finkelstein et al. 2015), including old or dusty galaxies. Contaminants for high-redshift samples constructed from color-color

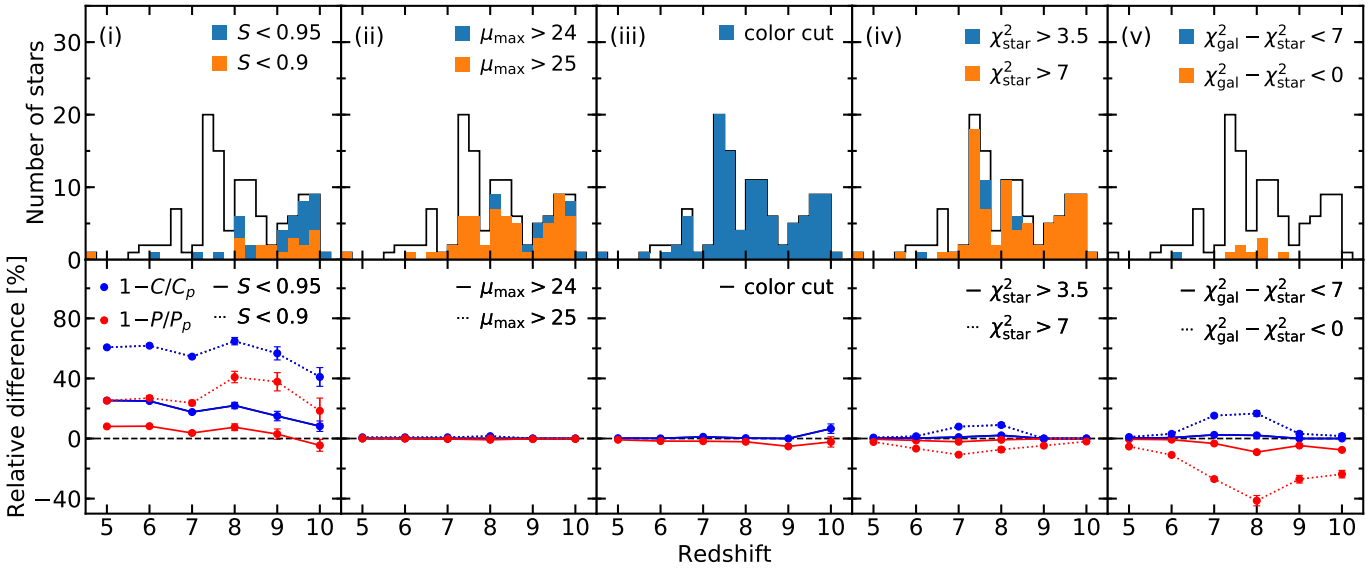

Fig. 17. *Top panels:* number of remaining stellar contaminants after applying the indicated selection criterion per photometric redshift interval, in the CEERS_1 observing strategy. Each column corresponds to one type of selection from Sect. 5.1. The black line indicates the photometric redshift distribution of the detected stars with $z_{\rm phot} > 5$. All the stars with lower photometric redshifts have $z_{\rm phot} < 2$. The colored bars represent the remaining stars after applying the indicated selection criterion. The soft selection (in blue) is always less restrictive than the stringent one (in orange), meaning that all second counts are also included in the first counts. *Bottom panels:* completeness C (in blue) and purity P (in red) of the high-redshift galaxy samples versus true redshift (integrated over magnitudes). The references C_p and P_p represent the completeness and purity assuming the selection is based on photometric redshifts only. The relative difference with respect to the reference is represented, so that positive values indicate lower completeness and purity. Different line styles represent the results for different selection criteria (as indicated in each panel).

Table 3. Sets of criteria to select high-redshift galaxies.

Set	Field		Criteria
Bouwens-like	EGS		$((m_{F606W} - m_{F814W} > 1.0) \wedge (m_{F115W} - m_{F150W} < 0.5) \wedge (m_{F606W} - m_{F814W} > 2.2(m_{F115W} - m_{F150W}) + 1.2)$
		$\vee$	$(m_{F814W} - m_{F115W} > 1.1) \wedge (m_{F115W} - m_{F150W} < 0.6) \wedge (m_{F814W} - m_{F115W} > 1.1(m_{F115W} - m_{F150W}) + 1.4)$
		$\vee$	$(m_{F115W} - m_{F150W} > 1.0) \wedge (m_{F150W} - m_{F200W} < 0.7) \wedge (m_{F115W} - m_{F150W} > 0.8(m_{F150W} - m_{F200W}) + 1.0))$
		$\wedge$	$z_{\rm phot}$ in z_i interval
	XDF		$((m_{F606W} - m_{F775W} > 0.8) \wedge (m_{F115W} - m_{F150W} < 0.8) \wedge (m_{F606W} - m_{F775W} > 1.5(m_{F115W} - m_{F150W}) + 1.0)$
		$\vee$	$(m_{F775W} - m_{F090W} > 0.6) \wedge (m_{F115W} - m_{F150W} < 0.8) \wedge (m_{F775W} - m_{F090W} > 0.8(m_{F115W} - m_{F150W}) + 0.8)$
		$\vee$	$(m_{F090W} - m_{F115W} > 0.7) \wedge (m_{F115W} - m_{F150W} < 1.0) \wedge (m_{F090W} - m_{F115W} > 0.7(m_{F115W} - m_{F150W}) + 0.9)$
		$\vee$	$(m_{F115W} - m_{F150W} > 0.8) \wedge (m_{F150W} - m_{F200W} < 1.0) \wedge (m_{F115W} - m_{F150W} > 0.8(m_{F150W} - m_{F200W}) + 0.9)$
		$\vee$	$(m_{F150W} - m_{F200W} > 0.8) \wedge (m_{F200W} - m_{F277W} < 1.0) \wedge (m_{F150W} - m_{F200W} > 0.8(m_{F200W} - m_{F277W}) + 0.9))$
		$\wedge$	$z_{\rm phot}$ in z_i interval
Bowler-like	all		$z_{\rm phot}$ in z_i interval
		$\wedge$	$((z_{\rm phot,sec} > z_i - \Delta z) \vee (\chi^2_{\rm sec} - \chi^2_{\rm gal} > 4))$
Finkelstein-like	all		PDFz integral under primary peak ≥ 0.7
		$\wedge$	PDFz integral in z_i interval ≥ 0.25
		$\wedge$	PDFz integral in z_i interval highest among intervals
		$\wedge$	PDFz integral in $[z_i - 1, \infty) \geq 0.5$
		$\wedge$	$(z_{\rm phot} > z_i - 2)$

Notes. The symbols $\wedge$ and $\vee$ represent the logical AND and OR, respectively.

criteria are usually low-redshift very red dusty galaxies or AGNs, and cool galactic stars. In the HUDF field, the deep HST optical imaging allows us to develop more redshift-specific color criteria. Nonetheless, we still rely on photometric redshift confirmation for these sources, especially at $z > 7$ where the NIRCam broad bands cannot precisely locate the Lyα break. The color criteria for the HUDF field are presented in Appendix B.

Alternatively, Bowler et al. (2015) criteria mainly use photometric redshifts and impose additional constraints on the location of the secondary photometric redshift $z_{\rm phot,sec}$. Similarly, Finkelstein et al. (2015) criteria make use of the whole posterior information to select objects based on the location and concentration of the PDFz in redshift intervals. In these two approaches, we do not include the criteria on the absolute quality of the

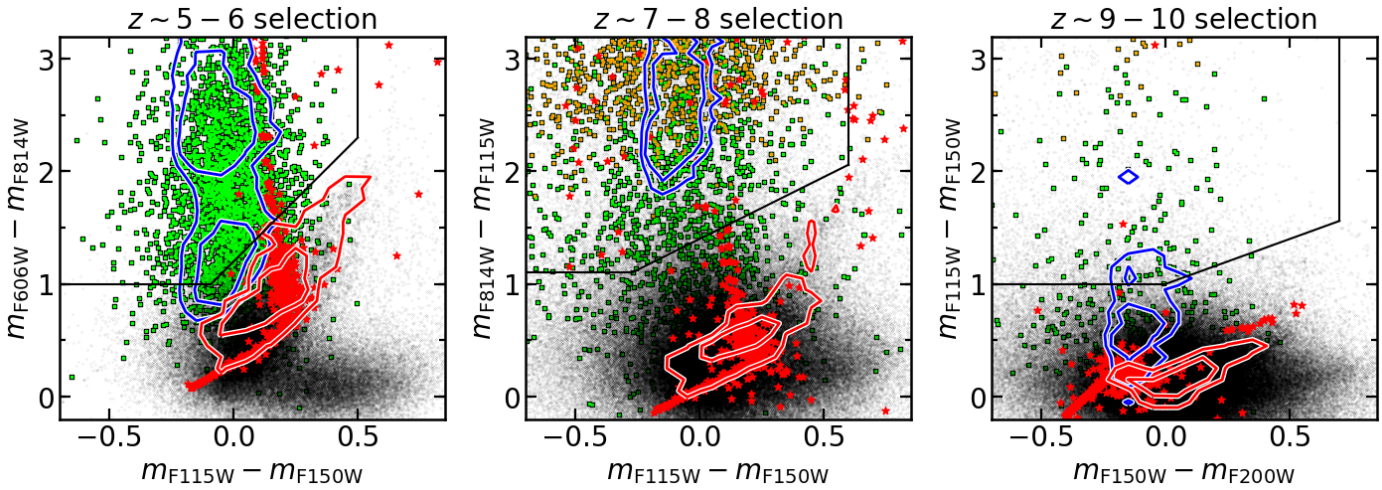

Fig. 18. Color-color selection criteria to preselect galaxies at $z \sim 5$–6, $z \sim 7$–8, $z \sim 9$–10 in the EGS field, with the Bouwens-like criteria in Table 3. The regions enclosed by the solid black line in the top-left corners show the color-color space region in which galaxies are preselected. The blue contours enclose 50% and 80% of the $z > 4.5$, $z > 6.5$, $z > 8.5$ galaxies input colors (without photometric scatter), and the red contours represent low-redshift quiescent galaxies. Each marker represents the measured colors for a detected source in the CEERS_1 observing strategy. Only sources detected at 5σ in the 3, 2, 2 reddest bands, respectively, are indicated. The green and orange squares are $z > 4.5$, $z > 6.5$, $z > 8.5$ galaxies, the orange squares indicating 1σ upper limits in the bluest band. The red stars are stellar objects, the black dots are low-redshift contaminants.

galaxy templates fit. Such criteria generally have a marginal impact on the final selection and, in our simulation, may just capture the differences between the input and the fitted templates.

Figure 19 illustrates the completeness and purity of the high-redshift galaxy samples in the CEERS_1 strategy, as a function of apparent observed-frame UV magnitudes m_{UV}. The colored lines represent the three different selections. The results for photometric redshift only selected sources (dotted lines), and the completeness of the detected sources assuming that redshifts are perfectly recovered (dashed lines), are also shown for comparison. The results for the CEERS_2 strategy are very similar.

We find that about $5 \pm 2\%$ ($2 \pm 1\%$) of the bright galaxies at $z = 4 - 6$ (1–2) are not detected. This implies that bright nearby objects contaminate the photometry of these sources, for which the source detection or the deblending procedure failed. At fainter magnitudes, the drop of completeness is the consequence of both this effect becoming stronger for faint sources and the impact of noise. Figure 20 illustrates the probability of finding a brighter neighbor in the input source catalog centered within the $0.5''$ diameter aperture. In the NIRCam/$F200W$ band, this probability is about 3% at 28 mag and converges to 13% at 33 mag. This gives a hint of the impact of blending alone on faint source photometry. Other scenarios are also possible, such as brighter neighbors outside of the aperture dominating the surface brightness of the faint source, therefore undetected or undeblended.

We find that the high-redshift galaxies selected through photometric redshifts only (before applying any other selection) already present significant incompleteness, even at $m_{UV} < 27$. Many sources which are correctly identified as high-redshift galaxies present relatively broad PDFz, so the resulting photometric redshifts often reside in the previous or next redshift interval. This is emphasized by the redshift intervals whose widths are fixed and not increasing with redshift. The bright high-redshift galaxies with catastrophic photometric redshifts are typically identified as red low-redshift galaxies, however many of them present PDFz with multiple peaks and a correct secondary solution. These results reflect the lack of deep optical and/or near-IR medium-band imaging, in the rest-frame UV region of these sources, to better identify the Lyα break, and the lack of blue-band imaging to confirm the break. Detected sources with nearby bright extended objects may also present contaminated photometry and colors, even for relatively bright galaxies. At very high redshift $z \geq 9$, the rarity of the galaxies of interest compared to the significantly more numerous low-redshift contaminants (at $z \sim 2$) leads to relatively low purity, in addition to the degeneracy between the Lyman break and the Balmer break at low redshift.

We observe slight differences between the different selection sets with respect to galaxy completeness and purity. Firstly, the Bouwens-like criteria lead to an improvement in purity, especially at bright magnitudes, with a relatively limited loss of completeness. Photometric redshifts impose most of the constraints, therefore the results are robust against small changes in the color preselection. Nonetheless, this preselection effectively increases purity, especially at the bright ends. Secondly, the Bowler-like selection induces a smaller loss of galaxy completeness, and increases the purity at the faint end. The criterion on the second peak of the PDFz has a significant impact on both C and P, especially at the faint ends. Thirdly, the criteria from Finkelstein lead to the highest galaxy completeness at most magnitudes and redshifts. At the same time, the resulting purity is the highest at the faint ends and at all redshifts, especially at $z < 8$. The constraint on the weight of the primary PDFz peak increases the purity and slightly decreases the completeness at the faint end. All the additional criteria increase even more the faint-end purity, however lowering the completeness at bright magnitudes. With these criteria, we find that the galaxy completeness is higher than 50% for $m_{UV} < 27.5$ sources at all redshifts, and purity remains above 80 and 60% at $z \leq 7$ and 10, respectively. From this comparison, we conclude that the PDFz criteria of Finkelstein result in the best trade-off between completeness and purity, and we keep these criteria in the next sections.

Figure B.2 illustrates the same analysis in the HUDF_1 configuration. The completeness after selecting galaxies is much

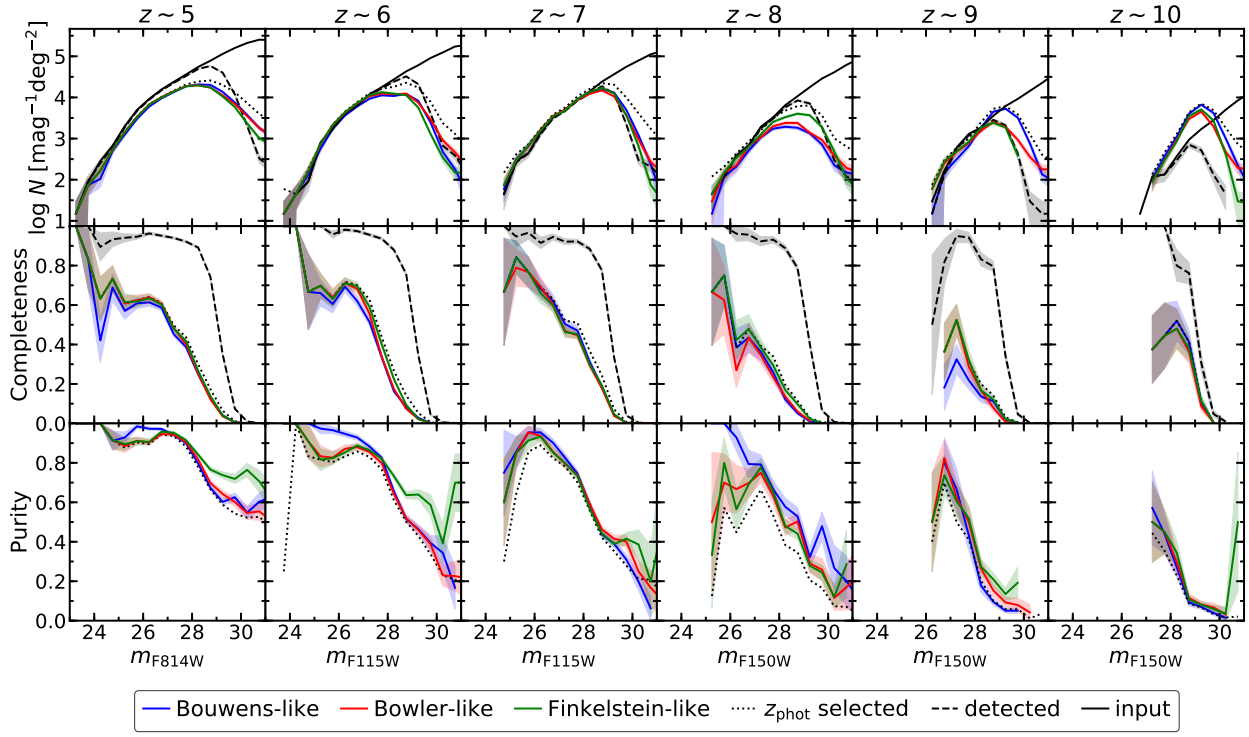

Fig. 19. High-redshift galaxy number counts (*top panels*), completeness (*middle panels*) and purity (*bottom panels*) versus apparent magnitude (in the band the nearest to rest-frame UV), in the CEERS_1 observing strategy. Each column corresponds to one redshift interval. Each colored line represents one set of selection criteria from Table 3. The solid black lines indicate the input number counts. The dashed black lines illustrate detected sources assuming the redshifts are perfectly recovered. The dotted black lines represent observed sources selected with photometric redshifts only. The shaded areas correspond to 1σ errors. The input and detected counts, and the measured completeness, are expressed in true magnitudes, while the selected counts and purity measurements are in observed magnitudes.

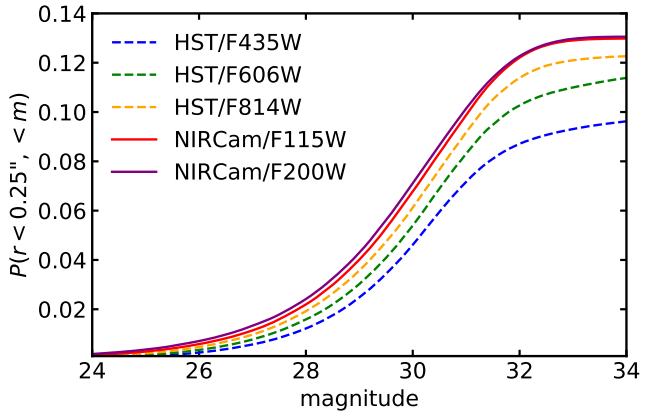

Fig. 20. Probability of finding at least one brighter neighbor in the input catalog centered within a 0.25″ radius. The line styles and colors represent different bands.

closer to the completeness assuming perfectly recovered redshifts. This is mainly due to the deeper NIRCam imaging and the additional HST *B* band. We observe similar features between the three selection sets as for the CEERS configuration. With the Finkelstein selection, the galaxy completeness remains higher than 50% at $m_{\rm UV} < 29$ at all redshifts, and the purity above 80% at $m_{\rm UV} < 30$.

Furthermore, completeness and purity may a priori depend on other physical parameters such as galaxy size. Completeness is about twice larger for galaxies with input effective radius $r_{\rm e} < 0.2$ kpc at $m_{\rm UV} > 29$ in CEERS and $m_{\rm UV} > 31$ in the HUDF. These sources are right at the detection limits, where completeness is only a few percent. We observe no significant evidence of purity varying with galaxy size. This variability should be taken into account when computing the luminosity function. Nonetheless, previous studies (e.g., Grazian et al. 2012; Finkelstein et al. 2015) found that the luminosity functions derived with or without including the size-luminosity relation remain similar. Despite this statement may depend on the assumed galaxy morphology, especially at high redshift, we decide to neglect the galaxy size variability of the completeness in the next sections.

5.3. Number counts predictions

We quantify the number of detected and selected sources in the high-redshift galaxy samples. Figures 19 and B.2 show the predicted number counts per magnitude and redshift, for the CEERS_1 and HUDF_1 observing strategies. The results are equivalent for the CEERS_2 and HUDF_2 configurations, respectively. The selected counts designate the selected objects following the indicated selection and assigned to the corresponding redshift interval. These are computed using observed magnitudes. In contrast, the detected and the input counts are computed using true magnitudes and redshifts. The drop at $m_{\rm UV} > 31$ at all redshifts comes from the stellar mass lower limit in the input galaxy catalog. The apparent disagreement between the input and the selected counts at $z \sim 10$ comes from photometric scatter.

For the CEERS_1 observing strategy, we expect about 916, 435, 232, 56, 19, 7 true high-redshift galaxies at $m_{\rm UV} < 29$ which are correctly assigned to the selected samples at $z \sim 5, 6, 7, 8, 9, 10$, respectively. In comparison, the input number

counts are 3039, 1522, 774, 318, 101, 21. These numbers agree with the predictions from the CEERS program description ($20-80$ sources at $z = 9-13$), though closer to the lower bound. One explanation may be source blending and the resulting increase in the photometric redshift outlier rates. Faint sources may even not be detected because of bright nearby objects, especially bright extended galaxies and stars, lowering the detected number counts. In addition, the high-redshift number counts importantly depend on the assumed evolution of the UVLF at $z > 8$, so that the rapid evolution assumed here gives lower number counts than with a slower evolution. For the HUDF_1 configuration, we expect 205, 135, 65, 20, 6, 2 selected sources at $m_{\rm UV} < 31$ and $z \sim 5, 6, 7, 8, 9, 10$, respectively, compared to the 628, 367, 222, 112, 40, 12 input counts. These numbers indicate that the GTO programs in the HUDF are more suitable than CEERS to study very faint galaxies at $z \geq 8$, in which case deeper imaging is required. On the other hand, the larger survey area in the EGS field enables more galaxies at $z \sim 5{-}6$ to be detected, including rare intrinsically bright sources.

5.4. Computing the galaxy luminosity function

In this section, we discuss the computation of the galaxy UV luminosity function from the selected galaxy number counts and the measured completeness and purity. The luminosity function is the comoving volume number density as a function of the intrinsic luminosity. The observed number density may suffer from incompleteness and impurities, therefore the observed LF needs to be corrected to recover the intrinsic LF using magnitude-dependent scaling factors.

The input galaxy UVLF in JAGUAR is constructed from the convolution of the stellar mass function and the $M_{\rm UV}(M_*)$ relation. Because of the stellar mass lower limit $\log(M_*/M_\odot) > 6$, the LF decreases at the faint end with a maximum situated between $-16 < M_{\rm UV} < -15$ at $4 < z < 10$. The position of this turn-over is still debated in the literature (e.g., Livermore et al. 2017; Bouwens et al. 2017), therefore we restrict ourselves to $M_{\rm UV} < -16$ where the faint end remains almost linear. We fit this input UVLF at $M_{\rm UV} > -22$ with a double-power-law model (DPL), parametrized as (Bowler et al. 2015):

$$\phi(M) = \frac{\phi^*}{10^{0.4(\alpha+1)(M-M^*)} + 10^{0.4(\beta+1)(M-M^*)}}, \quad (3)$$

where ϕ^* and M^* are the characteristic density and magnitude, α and β denote the faint and bright-end slopes. The difference between the input UVLF and the fitted model at $z \leq 10$ is at most 10% between $-22 < M_{\rm UV} < -17$.

We make forecasts for the recovery of the UVLF with the following approach. We take the selected galaxy $M_{\rm UV}$ number counts from our simulations, multiply them by the ratio of the survey area to the simulated area, and sample Poisson random vectors taking these values as the mean. The sampled counts are then corrected for incompleteness and impurities through the scaling correction factors, estimated from the number of input sources (function of true magnitudes) divided by the number of selected objects (function of observed magnitudes) from our simulations. This scaling therefore includes photometric scatter and $M_{\rm UV}$ recovery. We recall that absolute magnitudes are not corrected for dust attenuation. We use the classic estimator of the LF (Felten 1976), consisting of the absolute UV magnitude number counts divided by the comoving volume in the whole redshift interval. The LF uncertainties are the quadratic sum of the Poisson errors, cosmic variance errors (Trenti & Stiavelli 2008) and scaling correction uncertainties. By construction, the corrected LF values equal the input LF ones, however the uncertainties are broadened depending on the selected sample sizes. We fit each scaled Poisson random vector at $M_{\rm UV} < -16$ with a DPL model, using flat priors and a Markov chain Monte Carlo (MCMC) method to sample the posterior probability distribution. We finally marginalize over the Poisson samplings to determine the median parameters and errors. Both the statistical and the systematic errors on the parameters are included, though in reality one would have one Poisson sampling and only determine the statistical errors.

The recovered LFs are presented in Fig. 21 and Table A.1 for the CEERS_1 configuration covering about 100 arcmin^2. We do not present the results for the HUDF strategy, because the 4.7 arcmin^2 survey area cannot impose much statistical constraint on the LF, despite the increased depths. The differences between the selected and the corrected counts are significant, especially beyond $M_{\rm UV} > -18$ where the galaxy samples become highly incomplete. At $M_{\rm UV} < -18$, the scaling corrections are still $\sim$1–2 at all redshifts. Poisson uncertainties dominate the LF error budget at the bright end where the number counts are low, while cosmic variance errors reach up to 70% of the total variance at fainter magnitudes. In practice, large-scale structure effects will impact all magnitude the bins in a coherent way, but depend somewhat on the bias (e.g., Robertson et al. 2014). Scaling corrections contribute to about 10% of the total variance at almost all magnitudes and redshifts. As with real images, the corrections remain strongly dependent on the modeling assumptions, including galaxy morphology, star formation histories and dust attenuation. These results reflect that accurate simulations are required to correctly recover the galaxy counts, which can be severely affected by incompleteness and contamination. The number counts brightward of M^* decrease with increasing redshift, leading to a lack of constraints on the bright end. For this reason, we fix the DPL parameters β and M^*, at $z \geq 9$, to the input values when performing the fit. The obtained parameters are presented in Table 4. The faint-end slopes are effectively constrained with an absolute error of $\sim$0.1 at $z \leq 7$ and $\sim$0.25 at $z \geq 8$.

Within the CEERS area of 100 arcmin^2, the input galaxy UVLF predicts about 71, 36, 19, 12, 3.3 and 1.3 input galaxies with $M_{\rm UV} < M^*$ at $z \sim 5, 6, 7, 8, 9, 10$, respectively. These numbers indicate that the bright end of the UVLF cannot be constrained at $z \geq 7$, even assuming that all these galaxies are identified. Nonetheless, the NIRCam GTO program in the GOODS fields covering 200 arcmin^2, particularly the "Medium" survey, will bring additional constraints on the bright end of the UVLF up to $z \leq 8$. In spite of the depths of these programs, the main limitation remains the small JWST field of view. As an alternative, the Euclid[16] deep fields will include optical and near-IR imaging extended over tens of square degrees (Laureijs et al. 2011). These surveys, with the optical (e.g., Subaru Hyper Suprime-Cam) and mid-infrared (e.g., *Spitzer* Legacy Survey) counterparts, will reach the required depth to identify high-redshift galaxies, despite a lower resolution than JWST. The Euclid Deep Fields will probe the bright end of the luminosity function up to $z \sim 7$ or more, which will provide constraints complementary to the deep JWST surveys.

In addition, we predict the recovery of the cosmic SFR density $\rho_{\rm SFR}$. We integrate the UVLF to $M_{\rm UV} = -16$. The UV luminosity densities are converted into SFR densities using $\kappa_{\rm UV} = 1.15 \times 10^{-28}\, M_\odot\, {\rm yr}^{-1}({\rm erg/s/Hz})^{-1}$, where a 0.1–100 $M_\odot$ Salpeter initial mass function and a constant SFR are assumed

[16] http://www.euclid-ec.org

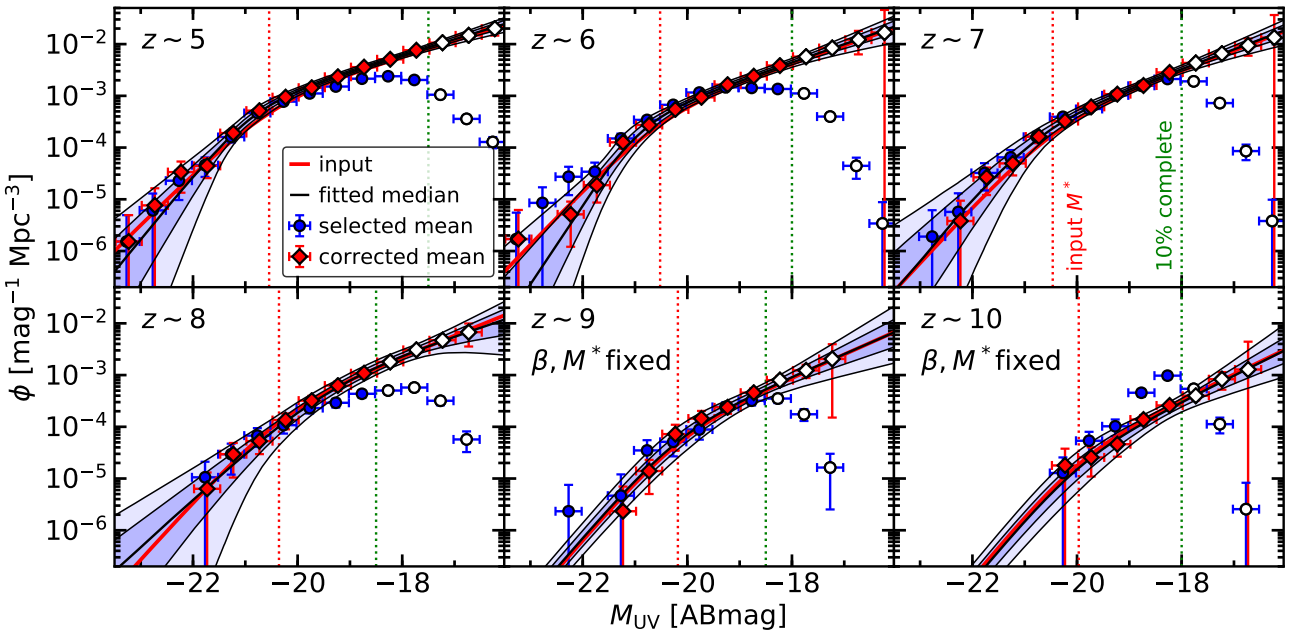

Fig. 21. Galaxy UV luminosity functions for multiple redshift intervals, for the CEERS_1 observing strategy. The blue dots indicate the estimated mean of the selected counts, and the red diamonds represent the selected counts corrected for incompleteness and impurity. The open symbols are the points where the completeness is below 10%. The error bars are Poisson errors for the former and the quadratic sum of Poisson and scaling errors for the latter. Multiple Poisson random vectors are sampled from the blue dots, scaled to correct for incompleteness and impurity and fitted with a double power-law function, with the indicated fixed parameters. The black lines show the model with the median fitted parameters, after marginalizing over all the sampled counts. The colored areas indicate 1σ and 2σ credibility errors. The red lines represent the input luminosity functions (fitted with a DPL model). The dotted red lines show M^* from the input LF, and the dotted green lines indicate the 10% completeness limit.

Table 4. Parametric fitting of the recovered UVLF.

z	ϕ^* [10^{-3} Mpc^{-3}]	M^* [mag]	α	β	$\log\rho_{\rm SFR}$ [$M_\odot$ yr^{-1} Mpc^{-3}]
			Input		
5	0.92	−20.54	−1.78	−3.50	−1.64
6	0.55	−20.52	−1.87	−3.63	−1.80
7	0.35	−20.46	−1.96	−3.73	−1.96
8	0.24	−20.36	−2.03	−3.79	−2.11
9	0.09	−20.18	−2.13	−3.95	−2.53
10	0.04	−19.97	−2.22	−4.07	−2.96
			Recovered		
5	$0.80^{+0.54}_{-0.27}$	$-20.84^{+0.39}_{-0.27}$	$-1.77^{+0.10}_{-0.08}$	$-4.10^{+0.71}_{-1.21}$	$-1.57^{+0.03}_{-0.03}$
6	$0.42^{+0.44}_{-0.17}$	$-20.85^{+0.48}_{-0.35}$	$-1.89^{+0.15}_{-0.11}$	$-4.79^{+1.00}_{-1.56}$	$-1.75^{+0.03}_{-0.03}$
7	$0.32^{+0.39}_{-0.17}$	$-20.67^{+0.55}_{-0.53}$	$-1.94^{+0.15}_{-0.11}$	$-3.93^{+0.73}_{-1.18}$	$-1.89^{+0.05}_{-0.04}$
8	$0.96^{+2.45}_{-0.76}$	$-19.39^{+1.04}_{-1.05}$	$-1.84^{+0.49}_{-0.28}$	$-3.27^{+0.43}_{-1.13}$	$-2.13^{+0.07}_{-0.07}$
9	$0.10^{+0.04}_{-0.03}$	−20.18	$-2.09^{+0.24}_{-0.22}$	−3.95	$-2.51^{+0.10}_{-0.10}$
10	$0.03^{+0.02}_{-0.01}$	−19.97	$-2.25^{+0.25}_{-0.27}$	−4.07	$-3.00^{+0.10}_{-0.09}$

(Madau & Dickinson 2014). The results, uncorrected for dust attenuation, are reported in Table 4. The SFR densities are correctly recovered and the expected errors remain below 0.1 dex, as long as the faint-end slope is well constrained. However, the errors are underestimated because of the fixed LF parameters at $z \geq 9$, and the scaling corrections recovering the input number counts. In addition, we do not apply any magnitude cuts, which would significantly lower the number of faint selected sources. In the ideal case where all the detected sources have perfectly recovered redshifts and absolute magnitudes, the errors on α and $\rho_{\rm SFR}$ are lowered by about 20% at $z < 8$. The cases at $z > 8$ are more sensitive to the determination of α from small number counts at the very faint end. Using all the input sources over the survey area at $\log(M_*/M_\odot) > 6$, we estimate that about 50% of the total errors arise from the limited area. This argues again in favor of surveys including larger cosmological volumes.

6. Summary and conclusion

In this paper, we forecast the performance of accepted JWST deep imaging surveys regarding the detection and analysis of high-redshift galaxies. In particular, we estimate the galaxy physical parameters, optimize the candidate selection with respect to galaxy completeness, purity and the total number of sources, then compute the UV luminosity function and the cosmic star formation rate density. We treat two JWST imaging programs, including CEERS in the EGS field, and HUDF GTO, and simulate the ancillary HST data for these fields. We construct complete mock samples of galaxies, local stars and brown dwarfs, representative of the current understanding of these populations using the latest observed luminosity and mass functions extrapolated to low masses, and high redshifts. The photometry of these sources is simulated through astronomical image generation, following the current knowledge of the JWST instruments. We extract the sources with SExtractor and estimate the source physical properties using SED-fitting.

Our main results can be summarized as follows:

– We find that the photometric redshifts estimated in the CEERS configuration are mainly limited by the lack of blue-band data. The additional MIRI bands marginally improve the photometric redshifts at faint magnitudes and at high redshift, where MIRI covers the rest-frame optical. Source blending contributes to up to 20% of the photometric redshift outliers in CEERS, and 40% in the HUDF.

– Stellar masses are recovered within 0.2 dex at $z \leq 5$ and 0.25 dex at $z > 5$, and are systematically overestimated by 0.1 dex at high redshift. Star formation rates are scattered over 0.3 dex and the most star-forming galaxies have a systematic bias of 0.1 to 0.2 dex. Numerous catastrophic SFR estimates arise from photometric redshift outliers.
– Galactic brown dwarfs contaminating the $z \geq 5$ galaxy samples can be effectively discarded, reaching a residual density of < 0.01 arcmin^{-2}. The impact on galaxy completeness remains minimal, although dependent on the assumed galaxy morphology.
– We find that the $5 < z < 10$ galaxy selection based on the redshift posterior probability distribution from SED-fitting gives the best compromise between completeness and purity. In the CEERS configuration, galaxy completeness remains above 50% at $m_{\rm UV} < 27.5$ and purity is higher than 80 and 60% at $z \leq 7$ and 10, respectively. In the HUDF strategy, the galaxy samples are more than 50% complete at $m_{\rm UV} < 29$ and 80% pure at $m_{\rm UV} < 30$ at all redshifts.
– We provide scaling correction factors for the selected galaxy number counts to recover the intrinsic number counts in the CEERS configuration. The values typically range from 1 to 2 at $M_{\rm UV} < -18$, but increase a lot at fainter magnitudes. This scaling is sensitive to the source modeling used as input, the source extraction and template fitting procedure, as well as the choice of ancillary data. Thus, the provided factors are strictly valid when using the same procedure presented here. However, our results show how crucial these types of calculations are to correctly recovering the luminosity function.
– The faint-end slope of the galaxy UV luminosity function in CEERS can be recovered with an error of ± 0.1 at $z = 5$ and ± 0.25 at $z = 10$, despite the significant dependence on the correction factors. We estimate that at least 300 arcmin^2 would be necessary to constrain the bright end up to $z = 8$.

We remind the reader that our forecasts are based on future JWST and existing HST imaging data, meaning that we neglect ancillary spectroscopy and ground-based imaging which may improve the results. In addition, the UVCANDELS program will enlarge the wavelength coverage in the EGS field, which may modestly improve the estimated photometric redshifts and the purity of the high-redshift galaxy samples.

In the future, we plan to include more realistic galaxy morphologies and use our simulations to fully exploit data from JWST imaging surveys. In addition, we plan to extend our simulations to the Euclid Deep Fields.

Acknowledgements. OI acknowledges the funding of the French Agence Nationale de la Recherche for the project "SAGACE". CCW acknowledges support from the National Science Foundation Astronomy and Astrophysics Fellowship grant AST-1701546. ECL acknowledges support from the ERC Advanced Grant 695671 "QUENCH". LC acknowledges support from the Spanish Ministry for Science and Innovation under grants ESP2017-83197 and MDM-2017-0737 "Unidad de Excelencia María de Maeztu – Centro de Astrobiología (CSIC-INTA)". JPP acknowledges the UK Science and Technology Facilities Council and the UK Space Agency for their support of the UK's JWST MIRI development activities. KIC acknowledges funding from the European Research Council through the award of the Consolidator Grant ID 681627-BUILDUP.

References

Aniano, G., Draine, B. T., Gordon, K. D., & Sandstrom, K. 2011, PASP, 123, 1218
Arnouts, S., Moscardini, L., Vanzella, E., et al. 2002, MNRAS, 329, 355
Baraffe, I., Homeier, D., Allard, F., & Chabrier, G. 2015, A&A, 577, A42
Bertin, E. 2009, Mem. Soc. Astron. It., 80, 422
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Bielby, R., Hudelot, P., McCracken, H. J., et al. 2012, A&A, 545, A23
Birkmann, S. M., Ferruit, P., Rawle, T., et al. 2016, Proc., SPIE, 9904
Bisigello, L., Caputi, K. I., Colina, L., et al. 2016, ApJS, 227, 19
Bisigello, L., Caputi, K. I., Colina, L., et al. 2017, ApJS, 231, 3
Bisigello, L., Caputi, K. I., Colina, L., et al. 2019, ApJS, 243, 27
Bixler, J. V., Bowyer, S., & Laget, M. 1991, A&A, 250, 370
Boucaud, A., Bocchio, M., Abergel, A., et al. 2016, A&A, 596, A63
Bouwens, R. J., Illingworth, G. D., Oesch, P. A., et al. 2015, ApJ, 803, 34
Bouwens, R. J., Oesch, P. A., Illingworth, G. D., Ellis, R. S., & Stefanon, M. 2017, ApJ, 843, 129
Bowler, R.A.A., Dunlop, J.S., McLure, R.J., et al. 2015, MNRAS, 452, 1817
Brammer, G. B., van Dokkum, P. G., Franx, M., et al. 2012, ApJS, 200, 13
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Burgasser, A. J. 2007, ApJ, 659, 655
Caballero, J. A., Burgasser, A. J., & Klement, R. 2008, A&A, 488, 181
Caffau, E., Ludwig, H.-G., Bonifacio, P., et al. 2010, A&A, 514, A92
Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ, 533, 682
Caputi, K. I., Ilbert, O., Laigle, C., et al. 2015, ApJ, 810, 73
Castelli, F., & Kurucz, R. L. 2004, A&A, 419, 725
Chabrier, G., Baraffe, I., Allard, F., & Hauschildt, P. 2000, ApJ, 542, 464
Charlot, S., & Fall, S. M. 2000, ApJ, 539, 718
Chen, B., Stoughton, C., Smith, J. A., et al. 2001, ApJ, 553, 184
Chevallard, J., & Charlot, S. 2016, MNRAS, 462, 1415
Cowley, W., Baugh, C., Cole, S., Frenk, C., & Lacey, C. 2018, MNRAS, 474, 2352
Curtis-Lake, E., McLure, R. J., Dunlop, J. S., et al. 2016, MNRAS, 457, 440
Davidzon, I., Ilbert, O., Laigle, C., et al. 2017, A&A, 605, A70
Dawson, W. A., Schneider, M. D., Tyson, J. A., & Jee, M. J. 2016, ApJ, 816, 11
Driver, S. P., Andrews, S. K., da Cunha, E., et al. 2018, MNRAS, 475, 2891
Dunlop, J. S. 2013, The First Galaxies (Astrophysics and Space Science Library), eds. T. Wiklind, B. Mobasher, & V. Bromm, 223
Felten, J. E. 1976, ApJ, 207, 700
Finkelstein, S. L., Ryan, R. E., Jr, Papovich, C., et al. 2015, ApJ, 810, 71
Fitzpatrick, E. L. 1999, PASP, 111, 63
Fitzpatrick, E. L., & Massa, D. 1986, ApJ, 307, 286
Furlanetto, S., Mirocha, J., Mebane, R., & Sun, G. 2017, MNRAS, 472, 1576
Galametz, A., Grazian, A., Fontana, A., et al. 2013, ApJS, 206, 10
Galliano, F., Hony, S., Bernard, J.-P., et al. 2011, A&A, 536, A88
Gardner, J. P., Mather, J. C., Clampin, M., et al. 2006, Proc., SPIE, 6265
Grazian, A., Castellano, M., Fontana, A., et al. 2012, A&A, 547, A51
Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, ApJS, 197, 35
Guo, Y., Ferguson, H. C., Giavalisco, M., et al. 2013, ApJS, 207, 24
Gutkin, J., Charlot, S., & Bruzual, G. 2016, MNRAS, 462, 1757
Hildebrandt, H., Erben, T., Kuijken, K., et al. 2012, MNRAS, 421, 2355
Hughes, D. H., Serjeant, S., Dunlop, J., et al. 1998, Nature, 394, 241
Ilbert, O., Tresse, L., Zucca, E., et al. 2005, A&A, 439, 863
Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841
Ilbert, O., Capak, P., Salvato, M., et al. 2009, ApJ, 690, 1236
Ilbert, O., McCracken, H. J., Le Fèvre, O., et al. 2013, A&A, 556, A55
Ilbert, O., Arnouts, S., Floc'h, E. L., et al. 2015, A&A, 579, A2
Illingworth, G. D., Magee, D., Oesch, P. A., et al. 2013, ApJS, 209, 6
Inoue, A. K., Shimizu, I., Iwata, I., & Tanaka, M. 2014, MNRAS, 442, 1805
Kemp, T. W., Dunlop, J. S., McLure, R. J., et al. 2019, MNRAS, 486, 3087
Koekemoer, A. M., Aussel, H., Calzetti, D., et al. 2007, ApJS, 172, 196
Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, ApJS, 197, 36
Krist, J. E., Hook, R. N., & Stoehr, F. 2011, Proc., SPIE, 8127
Kron, R. G. 1980, ApJS, 43, 305
Laigle, C., Davidzon, I., Ilbert, O., et al. 2019, MNRAS, 486, 5104
Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints [arXiv:1110.3193]
Le Fèvre, O., Tasca, L. A. M., Cassata, P., et al. 2015, A&A, 576, A79
Leauthaud, A., Massey, R., Kneib, J.-P., et al. 2007, ApJS, 172, 219
Livermore, R. C., Finkelstein, S. L., & Lotz, J. M. 2017, ApJ, 835, 113
Madau, P., & Dickinson, M. 2014, ARA&A, 52, 415
Mason, C. A., Trenti, M., & Treu, T. 2015a, ApJ, 813, 21
Mason, C. A., Treu, T., Schmidt, K. B., et al. 2015b, ApJ, 805, 79
McLeod, D. J., McLure, R. J., Dunlop, J. S., et al. 2015, MNRAS, 450, 3032
McLeod, D. J., McLure, R. J., & Dunlop, J. S. 2016, MNRAS, 459, 3812
McLure, R. J., Dunlop, J. S., Bowler, R. A. A., et al. 2013, MNRAS, 432, 2696
Momcheva, I. G., Brammer, G. B., van Dokkum, P. G., et al. 2016, ApJS, 225, 27
Morley, C. V., Fortney, J. J., Marley, M. S., et al. 2012, ApJ, 756, 172
Morley, C. V., Marley, M. S., Fortney, J. J., et al. 2014, ApJ, 787, 78
Mortlock, D. J., Patel, M., Warren, S. J., et al. 2012, MNRAS, 419, 390
Moutard, T., Arnouts, S., Ilbert, O., et al. 2016, A&A, 590, A102
Oesch, P. A., Bouwens, R. J., Illingworth, G. D., et al. 2014, ApJ, 786, 108

Oesch, P. A., Bouwens, R. J., Illingworth, G. D., Labbé, I., & Stefanon, M. 2018, ApJ, 855, 105
Oke, J. B. 1974, ApJS, 27, 21
Onodera, M., Renzini, A., Carollo, M., et al. 2012, ApJ, 755, 26
Pecaut, M. J., & Mamajek, E. E. 2013, ApJS, 208, 9
Pickles, A. J. 1998, PASP, 110, 863
Planck Collaboration VI. 2020, A&A, in press, https://doi.org/10.1051/0004-6361/201833910
Polletta, M., Tajer, M., Maraschi, L., et al. 2007, ApJ, 663, 81
Pontoppidan, K. M., Pickering, T. E., Laidler, V. G., et al. 2016, Proc., SPIE, 9910
Prevot, M. L., Lequeux, J., Maurice, E., Prevot, L., & Rocca-Volmerange, B. 1984, A&A, 132, 389
Ribeiro, B., Le Fèvre, O., Tasca, L. A. M., et al. 2016, A&A, 593, A22
Rieke, M. J., Kelly, D., & Horner, S. 2005, Proc., SPIE, 5904
Rieke, G. H., Wright, G. S., Böker, T., et al. 2015, PASP, 127, 584
Robertson, B. E., Ellis, R. S., Dunlop, J. S., et al. 2014, ApJ, 796, L27
Robertson, B. E., Ellis, R. S., Furlanetto, S. R., & Dunlop, J. S. 2015, ApJ, 802, L19
Robin, A. C., Reylé, C., Derriére, S., & Picaud, S. 2003, A&A, 409, 523
Robin, A. C., Marshall, D. J., Schultheis, M., & Reylé, C. 2012, A&A, 538, A106
Robin, A.C., Reylé, C., Fliri, J., et al. 2014, A&A, 569, A13
Salvato, M., Hasinger, G., Ilbert, O., et al. 2009, ApJ, 690, 1250
Sanders, D. B., Salvato, M., Aussel, H., et al. 2007, ApJS, 172, 86
Schenker, M. A., Robertson, B. E., Ellis, R. S., et al. 2013, ApJ, 768, 196
Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103
Schreiber, C., Pannella, M., Elbaz, D., et al. 2015, A&A, 575, A74
Schreiber, C., Elbaz, D., Pannella, M., et al. 2017, A&A, 602, A96
Schreiber, C., Elbaz, D., Pannella, M., et al. 2018, A&A, 609, A30
Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1
Shibuya, T., Ouchi, M., & Harikane, Y. 2015, ApJS, 219, 15
Sérsic, J. L. 1963, BAAA, 6, 41
Stefanon, M., Yan, H., Mobasher, B., et al. 2017, ApJS, 229, 32
Steidel, C. C., Giavalisco, M., Pettini, M., Dickinson, M., & Adelberger, K. L. 1996, ApJ, 462, L17
Tilvi, V., Papovich, C., Tran, K.-V. H., et al. 2013, ApJ, 768, 56
Tomczak, A. R., Quadri, R. F., Tran, K.-V. H., et al. 2014, ApJ, 783, 85
Trenti, M., & Stiavelli, M. 2008, ApJ, 676, 767
van der Wel, A., Bell, E. F., Häussler, B., et al. 2012, ApJS, 203, 24
van der Wel, A., Franx, M., van Dokkum, P. G., et al. 2014, ApJ, 788, 28
Wilkins, S. M., Bunker, A. J., Stanway, E., Lorenzoni, S., & Caruana, J. 2011, MNRAS, 417, 717
Williams, C. C., Curtis-Lake, E., Hainline, K. N., et al. 2018, ApJS, 236, 33
Wright, G. S., Wright, D., Goodson, G. B., et al. 2015, PASP, 127, 595
Wright, A. H., Driver, S. P., & Robotham, A. S. G. 2018, MNRAS, 480, 3491
Yung, L. Y. A., Somerville, R. S., Finkelstein, S. L., Popping, G., & Davé, R. 2019, MNRAS, 483, 2983
Zackrisson, E., Rydberg, C.-E., Schaerer, D., Östlin, G., & Tuli, M. 2011, ApJ, 740, 13

Appendix A: Additional table

Table A.1. Galaxy absolute magnitude number counts for luminosity function computation in CEERS_1 observing strategy.

$M_{\rm UV}$	$\mathbb{E}[N]$	C	P	S
		$z \sim 5$		
−22.75	0.8 ± 0.2	0.8 ± 0.2	1 ± 0	1.2 ± 0.3
−22.25	3.0 ± 0.3	0.7 ± 0.1	0.93 ± 0.06	1.5 ± 0.2
−21.75	6.0 ± 0.5	0.83 ± 0.07	0.83 ± 0.07	1.0 ± 0.1
−21.25	21.0 ± 0.9	0.76 ± 0.04	0.83 ± 0.04	1.19 ± 0.08
−20.75	61 ± 2	0.67 ± 0.03	0.75 ± 0.02	1.11 ± 0.05
−20.25	101 ± 2	0.64 ± 0.02	0.84 ± 0.02	1.22 ± 0.04
−19.75	146 ± 2	0.62 ± 0.02	0.85 ± 0.01	1.29 ± 0.03
−19.25	198 ± 3	0.52 ± 0.01	0.79 ± 0.01	1.54 ± 0.04
−18.75	281 ± 3	0.44 ± 0.01	0.73 ± 0.01	1.68 ± 0.04
−18.25	313 ± 4	0.283 ± 0.008	0.62 ± 0.01	2.13 ± 0.05
−17.75	266 ± 3	0.149 ± 0.005	0.48 ± 0.01	3.8 ± 0.1
−17.25	137 ± 2	0.044 ± 0.002	0.39 ± 0.02	10.2 ± 0.4
−16.75	47 ± 1	<0.01	0.42 ± 0.03	41 ± 3
−16.25	16.8 ± 0.8	<0.01	0.35 ± 0.05	155 ± 17
		$z \sim 6$		
−22.25	3.2 ± 0.4	1 ± 0	0.2 ± 0.1	0.2 ± 0.1
−21.75	4.0 ± 0.4	0.6 ± 0.1	0.6 ± 0.1	0.6 ± 0.1
−21.25	17.6 ± 0.8	0.70 ± 0.05	0.67 ± 0.05	0.84 ± 0.07
−20.75	40 ± 1	0.66 ± 0.04	0.63 ± 0.03	0.80 ± 0.05
−20.25	78 ± 2	0.68 ± 0.03	0.68 ± 0.02	0.82 ± 0.03
−19.75	136 ± 2	0.69 ± 0.02	0.62 ± 0.02	0.81 ± 0.03
−19.25	171 ± 3	0.53 ± 0.02	0.54 ± 0.02	1.12 ± 0.04
−18.75	166 ± 3	0.32 ± 0.01	0.47 ± 0.02	1.69 ± 0.06
−18.25	159 ± 3	0.169 ± 0.008	0.40 ± 0.02	2.8 ± 0.1
−17.75	130 ± 2	0.062 ± 0.004	0.33 ± 0.02	5.2 ± 0.2
−17.25	47 ± 1	<0.01	0.30 ± 0.03	21 ± 1
−16.75	5.2 ± 0.5	<0.01	0.35 ± 0.09	275 ± 54
−16.25	0.4 ± 0.1	<0.01	0.5 ± 0.4	4913 ± 3474
		$z \sim 7$		
−22.25	0.6 ± 0.2	1 ± 0	0.3 ± 0.3	0.7 ± 0.5
−21.75	3.4 ± 0.4	0.86 ± 0.09	0.88 ± 0.08	0.82 ± 0.05
−21.25	6.8 ± 0.5	0.85 ± 0.07	0.56 ± 0.09	0.8 ± 0.1
−20.75	16.8 ± 0.8	0.65 ± 0.05	0.65 ± 0.05	1.02 ± 0.09
−20.25	42 ± 1	0.60 ± 0.04	0.58 ± 0.03	0.82 ± 0.05
−19.75	58 ± 2	0.50 ± 0.03	0.51 ± 0.03	1.10 ± 0.06
−19.25	108 ± 2	0.44 ± 0.02	0.45 ± 0.02	1.05 ± 0.05
−18.75	164 ± 3	0.34 ± 0.02	0.32 ± 0.02	1.01 ± 0.04
−18.25	223 ± 3	0.22 ± 0.01	0.22 ± 0.01	1.33 ± 0.05

Notes. Per columns, the expectation value of the selected number counts, completeness, purity and scaling correction factor.

Table A.1. continued.

$M_{\rm UV}$	$\mathbb{E}[N]$	C	P	S
		$z \sim 7$		
−17.75	201 ± 3	0.063 ± 0.005	0.19 ± 0.01	2.27 ± 0.08
−17.25	76 ± 2	<0.01	0.21 ± 0.02	9.0 ± 0.5
−16.75	9.2 ± 0.6	<0.01	0.15 ± 0.05	108 ± 16
−16.25	0.4 ± 0.1	<0.01	<0.01	3512 ± 2483
		$z \sim 8$		
−21.75	1.0 ± 0.2	0.7 ± 0.3	0.6 ± 0.2	0.6 ± 0.2
−21.25	2.8 ± 0.3	0.6 ± 0.1	0.6 ± 0.1	1.0 ± 0.2
−20.75	6.4 ± 0.5	0.5 ± 0.1	0.44 ± 0.09	0.8 ± 0.1
−20.25	10.4 ± 0.6	0.45 ± 0.06	0.58 ± 0.07	1.2 ± 0.2
−19.75	21.8 ± 0.9	0.39 ± 0.04	0.53 ± 0.05	1.4 ± 0.1
−19.25	28 ± 1	0.23 ± 0.02	0.44 ± 0.04	2.1 ± 0.2
−18.75	41 ± 1	0.16 ± 0.02	0.26 ± 0.03	2.5 ± 0.2
−18.25	48 ± 1	0.060 ± 0.008	0.21 ± 0.03	3.6 ± 0.2
−17.75	54 ± 1	0.014 ± 0.003	0.14 ± 0.02	5.4 ± 0.3
−17.25	30 ± 1	<0.01	0.09 ± 0.02	15 ± 1
−16.75	5.4 ± 0.5	<0.01	0.07 ± 0.05	118 ± 23
−16.25	<0.01	<0.01	<0.01	<0.01
		$z \sim 9$		
−21.25	0.4 ± 0.1	<0.01	<0.01	0.5 ± 0.6
−20.75	3.0 ± 0.3	0.5 ± 0.2	0.5 ± 0.1	0.4 ± 0.1
−20.25	4.4 ± 0.4	0.52 ± 0.09	0.6 ± 0.1	1.4 ± 0.3
−19.75	7.6 ± 0.6	0.35 ± 0.06	0.37 ± 0.08	1.6 ± 0.3
−19.25	20.8 ± 0.9	0.27 ± 0.04	0.31 ± 0.05	1.0 ± 0.1
−18.75	27 ± 1	0.12 ± 0.02	0.16 ± 0.03	1.4 ± 0.1
−18.25	30 ± 1	0.05 ± 0.01	0.12 ± 0.03	2.3 ± 0.2
−17.75	15.0 ± 0.8	<0.01	0.11 ± 0.04	7.0 ± 0.9
−17.25	1.4 ± 0.2	<0.01	<0.01	126 ± 48
−16.75	<0.01	<0.01	<0.01	<0.01
−16.25	<0.01	<0.01	<0.01	<0.01
		$z \sim 10$		
−20.25	1.0 ± 0.2	0.3 ± 0.2	0.4 ± 0.2	1.4 ± 0.7
−19.75	4.2 ± 0.4	0.4 ± 0.2	0.3 ± 0.1	0.5 ± 0.1
−19.25	8.2 ± 0.6	0.6 ± 0.1	0.22 ± 0.06	0.4 ± 0.1
−18.75	36 ± 1	0.43 ± 0.07	0.08 ± 0.02	0.30 ± 0.04
−18.25	76 ± 2	0.16 ± 0.04	0.04 ± 0.01	0.26 ± 0.03
−17.75	42 ± 1	<0.01	0.03 ± 0.01	0.74 ± 0.08
−17.25	8.8 ± 0.6	<0.01	0.05 ± 0.03	7 ± 1
−16.75	0.20 ± 0.09	<0.01	<0.01	501 ± 501
−16.25	<0.01	<0.01	<0.01	<0.01

Appendix B: Galaxy selection at z > 5 in the HUDF

To select our high−redshift galaxies in the HUDF in the Bouwens-like set of criteria, we first preselect sources at $z \sim 5$ to 12 using the selection criteria in Table 3. These color criteria are represented in Fig. B.1. The high-redshift candidates are then confirmed with photometric redshifts. Figure B.2 indicates the high-redshift galaxy completeness and purity for the Bouwens-, Bowler- and Finkelstein-like criteria.

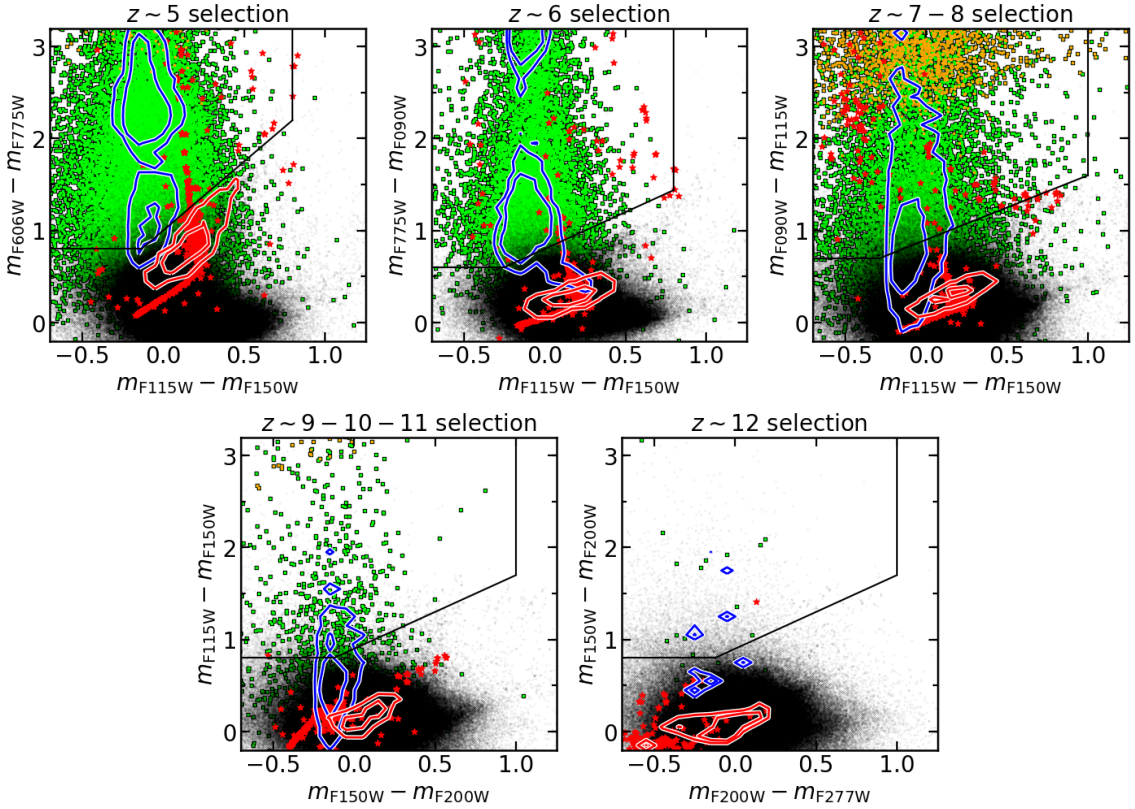

Fig. B.1. Same as Fig. 18, but in the HUDF, for the HUDF_1 observing strategy.

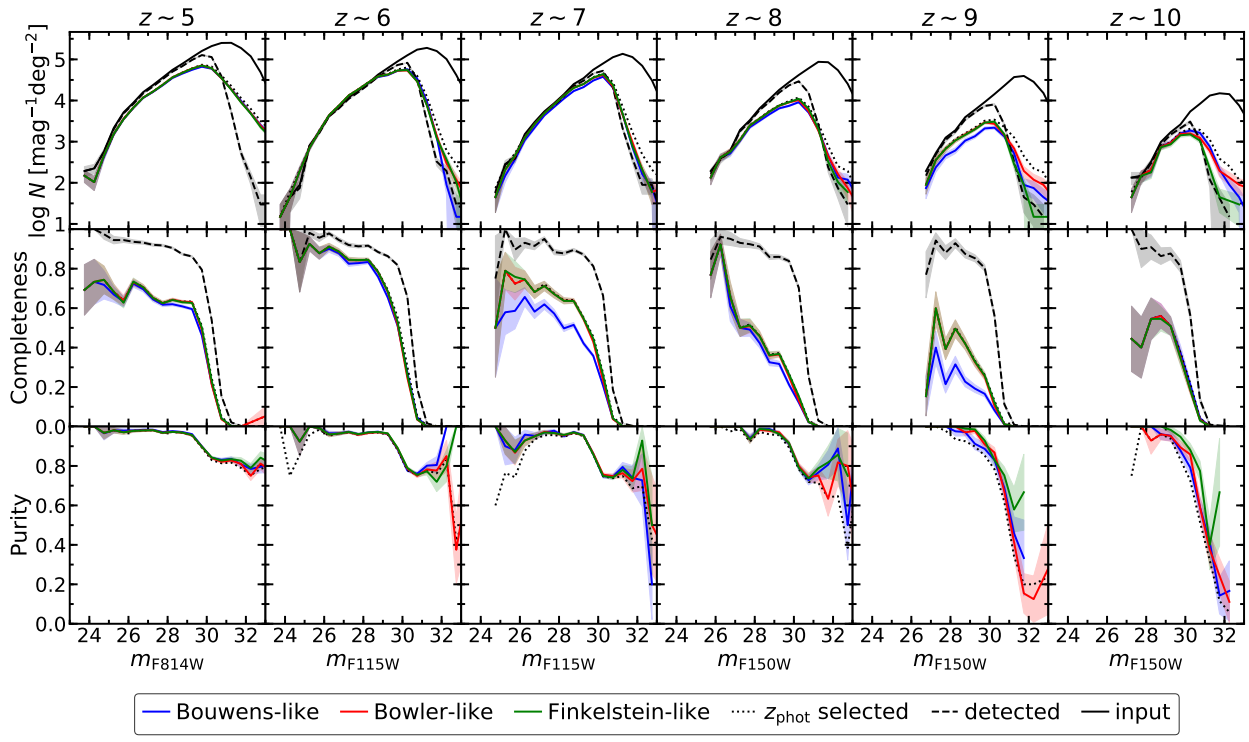

Fig. B.2. Same as Fig. 19, but for the HUDF_1 observing strategy.